The Correlated Pseudo-Marginal Method

George Deligiannidis[†], Arnaud Doucet[†] and Michael K. Pitt[‡*]

June 1, 2018

[†]University of Oxford, UK

[‡]King's College London, UK

Abstract

The pseudo-marginal algorithm is a Metropolis–Hastings-type scheme which samples asymptotically from a target probability density when we are only able to estimate unbiasedly an unnormalised version of it. In a Bayesian context, it is a state-of-the-art posterior simulation technique when the likelihood function is intractable but can be estimated unbiasedly using Monte Carlo samples. However, for the performance of this scheme not to degrade as the number T of data points increases, it is typically necessary for the number N of Monte Carlo samples to be proportional to T to control the relative variance of the likelihood ratio estimator appearing in the acceptance probability of this algorithm. The correlated pseudo-marginal method is a modification of the pseudo-marginal method using a likelihood ratio estimator computed using two correlated likelihood estimators. For random effects models, we show under regularity conditions that the parameters of this scheme can be selected such that the relative variance of this likelihood ratio estimator is controlled when N increases sublinearly with T and we provide guidelines on how to optimise the algorithm based on a non-standard weak convergence analysis. The efficiency of computations for Bayesian inference relative to the pseudo-marginal method empirically increases with T and exceeds two orders of magnitude in some examples.

Keywords: Asymptotic posterior normality; Correlated random numbers; Intractable likelihood; Metropolis–Hastings algorithm; Particle filter; Random effects model; Weak convergence.

1 Introduction

Consider a Bayesian model where the likelihood of the observations y is denoted by $p(y \mid \theta)$ and the prior for the parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ admits a density $p(\theta)$ with respect to Lebesgue measure $\mathrm{d}\theta$. Then the posterior density of interest is $\pi(\theta) \propto p(y \mid \theta)p(\theta)$. We slightly abuse notation by using the same symbols for distributions and densities.

A standard approach to compute expectations with respect to $\pi(\theta)$ is to use the Metropolis–Hastings (MH) algorithm to generate an ergodic Markov chain of invariant density $\pi(\theta)$. Given the current state θ of the Markov chain, one samples a candidate θ' which is accepted with a probability which depends in part on the likelihood ratio $p(y \mid \theta')/p(y \mid \theta)$. For many latent variable models, the likelihood is intractable and it is thus impossible to implement the MH algorithm. In this context, Markov chain Monte Carlo (MCMC) schemes targeting the joint posterior distribution of the parameter and latent variables are often inefficient as the parameter and latent variables can be strongly correlated under the posterior, or cannot even be used if only forward simulation of the latent variables is feasible; see, e.g., (Ionides et al., 2006), (Johndrow et al., 2016), and (Andrieu et al., 2010, Section 2.3) for a detailed discussion.

[*] *Address for correspondence*: M.K. Pitt, Department of Mathematics, King's College London, Strand, London, WC2R 2LS, UK. E-mail: michael.pitt@kcl.ac.uk

Contrary to these approaches, the pseudo-marginal (PM) algorithm directly mimics the MH scheme targeting the marginal $\pi(\theta)$ by substituting an estimator of the likelihood ratio $p(y \mid \theta')/p(y \mid \theta)$ for the true likelihood ratio in the MH acceptance probability (Lin et al., 2000; Beaumont, 2003; Andrieu and Roberts, 2009). This estimator is obtained by computing a non-negative unbiased estimator of $p(y \mid \theta')$ and dividing it by the estimator of $p(y \mid \theta)$ computed when θ was accepted. This simple yet powerful idea has become popular as it is often possible to obtain a non-negative unbiased estimator of intractable likelihoods and it provides state-of-the-art performance in many scenarios; see, e.g., (Andrieu et al., 2010) and (Flury and Shephard, 2011). Qualitative convergence results for this procedure have been obtained by Andrieu and Roberts (2009) and Andrieu and Vihola (2015).

Assuming that the likelihood estimator is evaluated using importance sampling or particle filters for state-space models with N particles, it has also been shown under various assumptions by Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015) that N should be selected such that the variance of the loglikelihood ratio estimator should take a value between 1.0 and 3.0 in regions of high probability mass to minimise the computational resources necessary to achieve a prespecified asymptotic variance for a particular PM average. As the number T of data $y = (y_1, ..., y_T)$ increases, this implies that N should increase linearly with T (Bérard et al., 2014, Theorem 1) and the computational cost of PM is thus of order T^2 at each iteration. This can be prohibitive for large datasets.

The reason for this is that the PM algorithm is based on an estimator of $p(y \mid \theta')/p(y \mid \theta)$ obtained by dividing estimators of $p(y \mid \theta)$ and $p(y \mid \theta')$ which are independent given θ and θ'. However, when one is interested in estimating a ratio, using positively correlated estimators of the numerator and denominator typically provides a lower variance ratio estimator than if these estimators were independent; see, e.g., Koop (1972). This is exploited by the proposed correlated pseudo-marginal (CPM) method which correlates these estimators by correlating the auxiliary random variates used to obtain them. Two implementations of this generic idea are detailed. We show how to correlate importance sampling estimators for random effects models and particle filter estimators for state-space models using the Hilbert sort procedure proposed by Gerber and Chopin (2015).

We study in detail the large sample properties of the CPM scheme for random effects models. In this scenario, the loglikelihood ratio estimator based on our correlation scheme satisfies a conditional Central Limit Theorem (CLT) whenever N grows to infinity sublinearly with T and the Euclidean distance between θ and θ' is of order $1/\sqrt{T}$. When the posterior concentrates towards a Gaussian density of standard deviation $1/\sqrt{T}$, this CLT can be used to show that a space-rescaled version of the CPM chain converges weakly to a discrete-time Markov chain on the parameter space. The Integrated Autocorrelation Time (IACT) of the weak limit is not impacted by how fast N goes to infinity with T. However the lower this growth rate is, the more correlated the auxiliary variables need to be to control the variance of this estimator. We provide results suggesting N needs to grow at least at rate $\sqrt{T}$ for the IACT of the original CPM chain to remain finite as $T \to \infty$. We use these results to provide practical guidelines on how to optimise the performance of the algorithm for large data sets which are validated experimentally. In our numerical examples on random effects models and state-space models, the CPM method always outperforms the PM method and the improvement increases with T from 20 to 50 times when T is a few hundred to more than 100 times when T is a few thousand.

The rest of the paper is organised as follows. In Section 2, we introduce the CPM algorithm and detail its implementation for random effects and state-space models. In Section 3, we present various CLTs for the loglikelihood estimator and loglikelihood ratio estimators used by the PM and CPM methods. In Section 4, we exploit these results to analyse and optimize the CPM kernel in the large sample regime. We demonstrate experimentally the efficiency of this methodology in Section 5 and discuss various potential extensions in Section 6. All the proofs are given in the Supplementary Material. The numerical results have been generated using Ox version 4.0 (Doornik, 2007). The computer code to replicate the experiments is available on line[1].

[1]Link: https://github.com/mikepitt1969/correlated

2 Metropolis–Hastings and correlated pseudo-marginal schemes

2.1 Metropolis–Hastings algorithm

The transition kernel Q_{MH} of the MH algorithm targeting $\pi(\theta)$ using a proposal distribution $q(\theta, \mathrm{d}\theta') = q(\theta, \theta')\,\mathrm{d}\theta'$ is given by

$$Q_{\text{MH}}(\theta, \mathrm{d}\theta') = q(\theta, \mathrm{d}\theta')\,\alpha_{\text{MH}}(\theta, \theta') + \{1 - \varrho_{\text{MH}}(\theta)\}\,\delta_{\theta}(\mathrm{d}\theta'), \tag{1}$$

where

$$r_{\text{MH}}(\theta, \theta') = \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')} = \frac{p(y \mid \theta')p(\theta')\,q(\theta', \theta)}{p(y \mid \theta)p(\theta)\,q(\theta, \theta')}, \tag{2}$$

and

$$\alpha_{\text{MH}}(\theta, \theta') = \min\{1, r_{\text{MH}}(\theta, \theta')\}, \quad \varrho_{\text{MH}}(\theta) = \int q(\theta, \mathrm{d}\theta')\,\alpha_{\text{MH}}(\theta, \theta'). \tag{3}$$

Implementing this MH scheme requires being able to evaluate the likelihood ratio $p(y \mid \theta')/p(y \mid \theta)$ appearing in the expression of $r_{\text{MH}}(\theta, \theta')$. When it is not possible to evaluate this ratio exactly, this MH algorithm cannot be implemented.

2.2 The correlated pseudo-marginal algorithm

Assume $\widehat{p}(y \mid \theta, U)$ is a non-negative unbiased estimator of the intractable likelihood $p(y \mid \theta)$ when $U \sim m$. Here U corresponds to the $\mathcal{U}$-valued auxiliary random variables used to obtain the estimator. We assume that $m(\mathrm{d}u) = m(u)\,\mathrm{d}u$ and introduce the joint density $\overline{\pi}(\theta, u)$ on $\Theta \times \mathcal{U}$, where

$$\overline{\pi}(\theta, u) = \pi(\theta)m(u)\,\widehat{p}(y \mid \theta, u)/p(y \mid \theta). \tag{4}$$

As $\widehat{p}(y \mid \theta, U)$ is unbiased, $\overline{\pi}(\theta, u)$ admits $\pi(\theta)$ as the marginal density. The CPM algorithm is an MH scheme targeting (4) with proposal density $q(\theta, \mathrm{d}\theta')\,K(u, \mathrm{d}u')$ where K admits an m-reversible Markov transition density, that is

$$m(u)\,K(u, u') = m(u')\,K(u', u). \tag{5}$$

This yields the acceptance probability

$$\alpha_Q\{(\theta, u), (\theta', u')\} = \min\left\{1, r_{\text{MH}}(\theta, \theta')\frac{\widehat{p}(y \mid \theta', u')/p(y \mid \theta')}{\widehat{p}(y \mid \theta, u)/p(y \mid \theta)}\right\}. \tag{6}$$

The CPM algorithm admits $\overline{\pi}(\theta, u)$ as an invariant density by construction and its transition kernel Q is given by

$$Q\{(\theta, u), (\mathrm{d}\theta', \mathrm{d}u')\} = q(\theta, \mathrm{d}\theta')\,K(u, \mathrm{d}u')\,\alpha_Q\{(\theta, u), (\theta', u')\} + \{1 - \varrho_Q(\theta, u)\}\,\delta_{(\theta,u)}(\mathrm{d}\theta', \mathrm{d}u'), \tag{7}$$

where $1 - \varrho_Q(\theta, u)$ is the corresponding rejection probability. For $K(u, u') = m(u')$, we recover the PM scheme. Data-informed proposals such as the preconditioned Crank-Nicolson Langevin proposal of Cotter et al. (2013) and its extensions proposed by Titsias and Papaspiliopoulos (2016) could also be used to update the auxiliary random variates at the cost of more complex acceptance probabilities.

Let $\varphi(z; \mu, \Sigma)$ be the multivariate normal density of argument z, mean μ and covariance matrix Σ and let $X \sim \mathcal{N}(\mu, \Sigma)$ denote a sample from this distribution. Henceforth, we focus on the case where the likelihood estimator is computed using $M \geq 1$ standard normal random variables and the corresponding Crank-Nicolson proposal (Cotter et al., 2013) is used. Hence we have

$$m(u) = \varphi(u; 0_M, I_M) \text{ and } K_\rho(u, u') = \varphi\left(u'; \rho u, \left(1 - \rho^2\right) I_M\right), \tag{8}$$

where $\rho \in (-1, 1)$, 0_M is the $M \times 1$ vector with zero entries and I_M the $M \times M$ identity matrix. It is straightforward to check that K_ρ is $m-$reversible. There is no loss of generality to select m as a normal density since inversion techniques can be used to form any random variable of interest[2].

[2]For example, in Section 2.3.2, it is necessary to generate uniform random variates and these may be constructed as $\Phi(u_i)$ where u_i is a scalar element of u and Φ the cumulative distribution function of the standard normal.

The selection of m as a normal and K_ρ as a proposal is advantageous because K_ρ can be interpreted as a discretised Ornstein–Uhlenbeck process. This is key in establishing the main theoretical result of Section 3 whose proof is simplified by the use of Itô's lemma and Stein's lemma. This allows us to provide useful guidelines on how to optimise the parameters of the CPM. Moreover, K_ρ is cheap to simulate from and admits a single interpretable parameter.

Algorithm 1 summarizes how to simulate from $Q\{(\theta, U), \cdot\}$.

Algorithm 1 Correlated Pseudo-Marginal Algorithm

1. Sample $\theta' \sim q(\theta, \cdot)$.
2. Sample $\varepsilon \sim \mathcal{N}(0_M, I_M)$ and set $U' = \rho U + \sqrt{1-\rho^2}\varepsilon$.
3. Compute the estimator $\widehat{p}(y \mid \theta', U')$ of $p(y \mid \theta')$.
4. With probability

$$\alpha_Q\{(\theta, U), (\theta', U')\} = \min\left\{1, \frac{\widehat{p}(y \mid \theta', U')}{\widehat{p}(y \mid \theta, U)} \frac{p(\theta')}{p(\theta)} \frac{q(\theta', \theta)}{q(\theta, \theta')}\right\}, \tag{9}$$

output (θ', U'). Otherwise, output (θ, U).

Contrary to the PM method corresponding to $\rho = 0$, we need to store the vector u instead of $\widehat{p}(y \mid \theta, u)$ to implement the algorithm when $\rho \neq 0$. In the applications considered, this overhead is mild.

The rationale behind the CPM scheme is that if $(\theta, u) \longmapsto \widehat{p}(y \mid \theta, u)$ is a regular enough function and (θ, U) and (θ', U') are "close" enough then we expect the ratio estimator $\widehat{p}(y \mid \theta', U')/\widehat{p}(y \mid \theta, U)$ to have small relative variance and therefore to better mimic the "exact" MH scheme Q_{MH}. In many situations, the posterior $\pi(\theta)$ will be approximately normal for large data sets with covariance scaling like $1/\sqrt{T}$, so an appropriately scaled MH random walk or autoregressive proposal $q(\theta, \mathrm{d}\theta')$ will ensure that θ and θ' are "close". We explain in Section 3 how ρ can be selected as a function of T to ensure that U and U' are "close" enough so that the loglikelihood ratio estimator $\log\{\widehat{p}(y \mid \theta', U')/\widehat{p}(y \mid \theta, U)\}$ satisfies a conditional CLT at stationarity. As explained in the introduction, properties of this estimator and in particular its asymptotic distribution and variance at stationarity are critical to our analysis of the CPM scheme in the large sample regime detailed in Section 4.

2.3 Application to latent variable models

2.3.1 Random effects models

Consider the model

$$X_t \overset{\text{i.i.d.}}{\sim} f_\theta(\cdot), \qquad Y_t \mid X_t \sim g_\theta(\cdot \mid X_t), \tag{10}$$

where $\{X_t; t \geq 1\}$ are $\mathbb{R}^k$-valued latent variables, $\{Y_t; t \geq 1\}$ are Y-valued observations, Y being a topological space, and $f_\theta(\cdot), g_\theta(\cdot \mid x)$ are densities with respect to the corresponding Lebesgue measure. For any $i < j$, let $i:j = \{i, i+1, ..., j\}$. For a realization $Y_{1:T} = y_{1:T}$, the likelihood satisfies

$$p(y_{1:T} \mid \theta) = \prod_{t=1}^{T} p(y_t \mid \theta), \qquad p(y_t \mid \theta) = \int g_\theta(y_t \mid x_t)\, f_\theta(x_t) \mathrm{d}x_t. \tag{11}$$

If the T integrals appearing in (11) are intractable, we can estimate them using importance sampling to obtain the following unbiased likelihood estimator

$$\widehat{p}(y_{1:T} \mid \theta, U) = \prod_{t=1}^{T} \left\{\frac{1}{N} \sum_{i=1}^{N} \omega(y_t, X_{t,i}; \theta)\right\}, \tag{12}$$

where the importance weight $\omega(y, U_{t,i}; \theta)$ is given by

$$\omega(y_t, U_{t,i}; \theta) = \frac{g_\theta(y_t \mid X_{t,i})\, f_\theta(X_{t,i})}{q_\theta(X_{t,i} \mid y_t)}, \tag{13}$$

assuming that there exists a deterministic map $\Xi_t : \mathbb{R}^p \times \Theta \to \mathbb{R}^k$ such that $X_{t,i} = \Xi_t(U_{t,i}; \theta) \sim q_\theta(\cdot \mid y_t)$ for $U_{t,i} \sim \mathcal{N}(0_p, I_p)$. Let U be the column vector consisting of all the components of $U_{t,i}$ for $t \in 1:T$ and $i \in 1:N$. It is clear that $U \sim \mathcal{N}(0_M, I_M)$ where $M = TNp$.

2.3.2 State-space models

Consider a generalization of the model (10) where the latent variables $\{X_t; t \geq 1\}$ now arise from a homogeneous $\mathbb{R}^k$-valued Markov process of initial density ν_θ and Markov transition density f_θ with respect to Lebesgue measure, i.e., for $t \geq 1$

$$X_1 \sim \nu_\theta, \quad X_{t+1} | X_t \sim f_\theta(\cdot \mid X_t), \qquad Y_t | X_t \sim g_\theta(\cdot \mid X_t). \tag{14}$$

For a realization $Y_{1:T} = y_{1:T}$, the likelihood satisfies the predictive decomposition

$$p(y_{1:T} \mid \theta) = p(y_1 \mid \theta) \prod_{t=2}^{T} p(y_t \mid y_{1:t-1}, \theta), \tag{15}$$

with

$$p(y_t \mid y_{1:t-1}, \theta) = \int g_\theta(y_t \mid x_t)\, p_\theta(x_t \mid y_{1:t-1}) \mathrm{d}x_t, \tag{16}$$

where $p_\theta(x_1 \mid y_{1:0}) = \nu_\theta(x_1)$ and $p_\theta(x_t \mid y_{1:t-1})$ denotes the posterior density of X_t given $Y_{1:t-1} = y_{1:t-1}$ for $t \geq 2$. Importance sampling estimators of the likelihood have relative variance typically increasing exponentially with T so the likelihood is usually estimated using particle filters instead.

Particle filters propagate N random samples, termed particles, over time using a sequence of resampling steps and importance sampling steps using the importance densities $q_\theta(x_1 | y_1)$ at time 1 and $q_\theta(x_t | y_t, x_{t-1})$ at times $t \geq 2$. Let $\Xi_1 : \mathbb{R}^p \times \Theta \to \mathbb{R}^k$ and $\Xi_t : \mathbb{R}^k \times \mathbb{R}^p \times \Theta \to \mathbb{R}^k$ for $t \geq 2$ be deterministic maps such that $X_1 = \Xi_1(V; \theta) \sim q_\theta(\cdot \mid y_1)$ and $X_t = \Xi_t(x_{t-1}, V; \theta) \sim q_\theta(\cdot \mid y_t, x_{t-1})$ for $t \geq 2$ if $V \sim \mathcal{N}(0_p, I_p)$. We also propose to use normal random variables to obtain the uniform random variables necessary to sample the categorical distributions appearing in the resampling steps. By using these representations, we obtain an unbiased estimator $\widehat{p}(y_t \mid \theta, U)$ of $p(y_t \mid \theta)$ where U follows a multivariate normal distribution (Del Moral, 2004). When this estimator is used within a PM scheme, the resulting algorithm is known as the particle marginal MH (Andrieu et al., 2010). However if this likelihood estimator is used in the CPM context, the likelihood ratio estimator $\widehat{p}(y_{1:T} \mid \theta', u')/\widehat{p}(y_{1:T} \mid \theta, u)$ can significantly deviate from 1 even when (θ, u) is close to (θ', u') and the true likelihood is continuous at θ. This is because the resampling steps introduce discontinuities in the particles that are selected when θ and u are modified, even slightly (Malik and Pitt, 2011).

To reduce the variability of this likelihood ratio estimator, we use a resampling scheme based on the Hilbert sort procedure introduced by Gerber and Chopin (2015). This procedure is based on the Hilbert space-filling curve which is a continuous fractal map $H : [0,1] \to [0,1]^k$ whose image is $[0,1]^k$. It admits a pseudo-inverse $h : [0,1]^k \to [0,1]$, that is $H \circ h(x) = x$ for all $x \in [0,1]^k$. For most points x, x' that are close in $[0,1]^k$, their images $h(x)$ and $h(x')$ tend to be close. This property can be used to build a "sorted" resampling procedure which will ensure that when the parameter or auxiliary variables change only slightly the particles that are selected remain close. Practically, this resampling procedure proceeds as follows: 1) the $\mathbb{R}^k-$valued particles are projected in the hypercube $[0,1]^k$ using a bijection $\varkappa : \mathbb{R}^k \to [0,1]^k$, 2) The resulting $[0,1]^k-$valued particles are projected on $[0,1]$ using the pseudo-inverse h, 3) These projected $[0,1]-$valued particles are sorted, 4) The systematic resampling scheme proposed by Carpenter et al. (1999) is used on the sorted points.

Let us introduce the importance weights $\omega_1(x_1; \theta) = \nu_\theta(x_1)\, g_\theta(y_1 \mid x_1)/\, q_\theta(x_1 \mid y_1)$ and $\omega_t(x_{t-1}, x_t; \theta) = f_\theta(x_t \mid x_{t-1})\, g_\theta(y_t \mid x_t)/q_\theta(x_t \mid y_t, x_{t-1})$ for $t \geq 2$. The only difference between the resulting particle filter presented below and the algorithm of Gerber and Chopin (2015) is that we use normal random variates instead of randomized quasi-Monte Carlo points in $[0,1]^p$. For the mapping $\varkappa$, we adopt the logistic transform used in Gerber and Chopin (2015).

Algorithm 2 Particle filter using Hilbert sort

1. Sample $U_{1,i} \sim \mathcal{N}(0_p, I_p)$ and set $X_{1,i} = \Xi_1(U_{1,i}; \theta)$ for $i \in 1:N$.
2. For $t = 1, \ldots, T-1$
   (a) Find the permutation σ_t such that $h \circ \varkappa\left(X_{t,\sigma_t(1)}\right) \leq \ldots \leq h \circ \varkappa\left(X_{t,\sigma_t(N)}\right)$ if $k \geq 2$, or $X_{t,\sigma_t(1)} \leq \ldots \leq X_{t,\sigma_t(N)}$ if $k = 1$.
   (b) Sample $U_t^R \sim \mathcal{N}(0,1)$, set $\overline{U}_{t,i} = (i-1)/N + \Phi(U_t^R)/N$ for $i \in 1:N$.
   (c) Sample $A_{t,i} \sim F_t^{-1}\left(\overline{U}_{t,i}\right)$ for $i \in 1:N$ where F_t^{-1} is the generalized inverse distribution function of the categorical distribution with weights $\{\omega_1(X_{1,\sigma_1(i)}; \theta); i \in 1:N\}$ if $t = 1$ and $\{\omega_t(X_{t-1,\sigma_{t-1}\left(A_{t-1,\sigma_t(i)}\right)}, X_{t,\sigma_t(i)}; \theta); i \in 1:N\}$ for $t \geq 2$.
   (d) Sample $U_{t+1,i} \sim \mathcal{N}(0_p, I_p)$ and set $X_{t+1,i} = \Xi_{t+1}(X_{t,\sigma_t(A_{t,i})}, U_{t+1,i}; \theta)$ for $i \in 1:N$.

If we denote by U the column vector composed of the components of $\left(U_{1,1}, ..., U_{T,N}, U_1^R, ..., U_{T-1}^R\right)$, then $U \sim \mathcal{N}(0_M, I_M)$ where $M = TNp + T - 1$. The corresponding unbiased likelihood estimator is given by

$$\widehat{p}(y_{1:T} \mid \theta, U) = \left\{ \frac{1}{N} \sum_{i=1}^{N} \omega_1(X_{1,i}; \theta) \right\} \prod_{t=2}^{T} \left\{ \frac{1}{N} \sum_{i=1}^{N} \omega_t(X_{t-1,\sigma_{t-1}(A_{t-1,i})}, X_{t,i}; \theta) \right\}. \tag{17}$$

We can now use this estimator within the CPM scheme. Many valid alternatives and generalizations of this scheme are possible as discussed in Section 6. For example, we found that introducing an additional Hilbert sort step after resampling can slightly improve performance without affecting the scaling properties.

2.4 Discussion

Ideas related to the CPM scheme have previously been proposed: Lee and Holmes (2010) suggest combining PM steps with updates where only θ is updated while U is held fixed, but this scheme scales poorly with T as it still uses PM steps. In (Andrieu et al., 2012), the authors propose combining PM steps with steps where θ is held fixed and correlation between $\widehat{p}(y \mid \theta, U)$ and $\widehat{p}(y \mid \theta, U')$ is introduced by sampling U' using an $m-$reversible Markov kernel K. However, the crucial selection of K was not discussed. It was independently proposed by Dahlin et al. (2015) to use the correlation scheme (8) but the guidelines for the correlation parameter ρ therein do not ensure that the variance of the loglikelihood ratio estimator is controlled as T increases. This work also relies on a standard particle filter.

As the density m of U is independent of θ, it might be argued that a Gibbs algorithm sampling alternately from the full conditional densities $\overline{\pi}(\theta \mid u)$ and $\overline{\pi}(u \mid \theta)$ of $\overline{\pi}(\theta, u)$ could mix well. Related ideas have been explored in (Papaspiliopoulos et al., 2007). Such a Gibbs strategy is usually not implementable in the applications considered here. Particle Gibbs samplers have been proposed to mimic this strategy but their computational complexity is of order T^2N per iteration for state-space models when using such a parameterisation (Lindsten et al., 2014, Section 6.2). Thus they are not competitive with the PM algorithm whose cost is of order T^2 per iteration. An alternative approach for updating U given θ, proposed by Murray and Graham (2016), is to use elliptical slice sampling. However, in this context, no guidelines for the selection of N have been proposed. Experimentally, this method is not competitive with an appropriately tuned CPM scheme when the same value of N is used for both methods. We observed that elliptical slice sampling is attempting many moves on the ellipse which are not on the support of the slice, thus requiring multiple expensive evaluations of the simulated likelihood for each sample.

3 Asymptotics of the loglikelihood ratio estimators

To understand the quantitative properties of the CPM scheme, it is key to establish the statistical properties of the likelihood ratio estimator appearing in its acceptance probability (6). For the random

effects models introduced in Section 2.3.1, we establish conditional CLTs for the loglikelihood estimator (12) and the corresponding loglikelihood ratio estimators used by the PM and the CPM algorithms when $N \to \infty$ and $T \to \infty$. Here N will be a deterministic function of T denoted by N_T. We show that these estimators exhibit very different behaviours, underlining the benefits of CPM over PM.

Consider a sequence of random variables $\{M^T; T \geq 1\}$ defined on a probability space $(\Omega, \mathcal{G}, P)$, a sequence of sub-σ-algebras $\{\mathcal{G}^T; T \geq 1\}$ and write $\to_P$ to denote convergence in probability. We also write $M^T \big| \mathcal{G}^T \Rightarrow \lambda$ if $M \sim \lambda$ and $\mathbb{E}[f(M^T) \big| \mathcal{G}^T] \to_P \mathbb{E}[f(M)]$ as $T \to \infty$ for any bounded continuous function f.

Henceforth, we will make the assumption that $Y_t \overset{\text{i.i.d.}}{\sim} \mu$ and write $\mathcal{Y}^T$ for the σ-field spanned by $Y_{1:T}$. When additionally $U \sim m$, we denote the associated probability measure, expectation and variance by $\mathbb{P}$, $\mathbb{E}$ and $\mathbb{V}$. As our limit theorems consider the asymptotic regime where $T \to \infty$ and $N_T \to \infty$, we should write m_T, π_T instead m, π and similarly U^T, U_t^T and $U_{t,i}^T$ instead of U, U_t and $U_{t,i}$. The probability space is defined precisely in Section A.1 of the Supplementary Material. We do not emphasise here this dependency on T for notational simplicity but it should be kept in mind that we are dealing with triangular arrays of random variables. We can write unambiguously $\mathbb{E}(\psi(Y_1, U_{1,1}; \theta))$ rather than $\mathbb{E}(\psi(Y_1, U_{1,1}^T; \theta))$ as $U_{1,1}^T \sim \mathcal{N}(0_p, I_p)$ under $\mathbb{P}$ for any $T \geq 1$.

3.1 Asymptotic distribution of the loglikelihood error

Let $\gamma(y_1; \theta)^2 = \mathbb{V}(\varpi(y_1, U_{1,1}; \theta))$ be the conditional variance given $Y_1 = y_1$ and $\gamma(\theta)^2 = \mathbb{V}(\varpi(Y_1, U_{1,1}; \theta)) = \mathbb{E}(\gamma(Y_1; \theta)^2)$ the unconditional variance of the normalized importance weight

$$\varpi(Y_t, U_{1,1}; \theta) = \frac{\omega(Y_t, U_{1,1}; \theta)}{p(Y_t \mid \theta)}, \tag{18}$$

where $\omega(Y_t, U_{1,1}; \theta)$ is defined in (13).

Our first result establishes conditional CLTs for the loglikelihood error

$$Z_T(\theta) = \log \widehat{p}(Y_{1:T} \mid \theta, U) - \log p(Y_{1:T} \mid \theta), \tag{19}$$

when U arises from the proposal m or from the equilibrium distribution

$$\overline{\pi}(u \mid \theta) = \frac{\overline{\pi}(\theta, u)}{\pi(\theta)} = \prod_{t=1}^{T} \frac{\widehat{p}(Y_t \mid \theta, u_t)}{p(Y_t \mid \theta)} \varphi(u_t; 0_{pN_T}, I_{pN_T}), \tag{20}$$

with $\overline{\pi}(\theta, u)$ as defined in (4).

Theorem 1. *Let $N_T = \lceil \beta T^\alpha \rceil$ with $1/3 < \alpha \leq 1$, $\beta > 0$ and $Y_t \overset{\text{i.i.d.}}{\sim} \mu$.*

1. *If $\mathbb{E}\left(\varpi(Y, U_{1,1}; \theta)^8\right) < \infty$ and $U \sim m$ then*

$$T^{(\alpha-1)/2} Z_T(\theta) + \frac{1}{2} T^{(1-\alpha)/2} \beta^{-1} \gamma(\theta)^2 \Big| \mathcal{Y}^T \Rightarrow \mathcal{N}\left(0, \beta^{-1} \gamma(\theta)^2\right). \tag{21}$$

2. *If $\mathbb{E}\left(\varpi(Y_1, U_{1,1}; \theta)^9\right) + \mathbb{E}\left(\gamma(Y_1; \theta)^4\right) < \infty$ and $U \sim \overline{\pi}(\cdot \mid \theta)$ then*

$$T^{(\alpha-1)/2} Z_T(\theta) - \frac{1}{2} T^{(1-\alpha)/2} \beta^{-1} \gamma(\theta)^2 \Big| \mathcal{Y}^T \Rightarrow \mathcal{N}\left(0, \beta^{-1} \gamma(\theta)^2\right). \tag{22}$$

Remark. To establish (21), respectively (22), for $1/2 < \alpha \leq 1$, the condition $\mathbb{E}\left(\varpi(Y_1, U_{1,1}; \theta)^4\right) < \infty$, respectively $\mathbb{E}\left(\varpi(Y_1, U_{1,1}; \theta)^5\right) < \infty$, is sufficient.

For particle filters, a CLT for $Z_T(\theta)$ of the form (21) has already been established for the case $\alpha = 1$ in (Bérard et al., 2014), when using multinomial resampling under strong mixing assumptions. We conjecture that both (21) and (22) hold under weaker assumptions for $1/3 < \alpha < 1$ and the Hilbert sort resampling scheme. However, it is very technically challenging to establish this result[3].

The result (21) suggests that for large T under the proposal, $Z_T(\theta)$ is approximately normal with mean $-\beta^{-1} T^{1-\alpha} \gamma(\theta)^2 / 2$ and variance $\beta^{-1} T^{1-\alpha} \gamma(\theta)^2$. The result (22) suggests that at equilibrium $Z_T(\theta)$ is approximately normal with the same variance but opposite mean.

[3] In the simpler scenario where one uses systematic resampling, such a CLT has not yet been established. Some of the technical problems which arise when attempting to carry out such an analysis are detailed in (Gentil and Rémillard, 2008).

3.2 Asymptotic distribution of the loglikelihood ratio error

Assume that we are at state (θ, U) and propose (θ', U') using $\theta' \sim q(\theta, \cdot)$, $U' \sim m$ as in the PM algorithm or $\theta' \sim q(\theta, \cdot)$, $U' \sim K_\rho(U, \cdot)$ as in the CPM algorithm. In both cases, the acceptance ratio (6) depends on the loglikelihood ratio error

$$R_T(\theta, \theta') = \log \frac{\widehat{p}(Y_{1:T} \mid \theta', U')}{\widehat{p}(Y_{1:T} \mid \theta, U)} - \log \frac{p(Y_{1:T} \mid \theta')}{p(Y_{1:T} \mid \theta)}. \tag{23}$$

We examine here the limiting distribution of $R_T(\theta, \theta + \xi/\sqrt{T})$ for fixed θ and ξ, the rationale being that the posterior typically concentrates at rate $1/\sqrt{T}$ when T increases. Thus a correctly scaled random walk proposal for an MH algorithm will be of the form $\theta' = \theta + \xi/\sqrt{T}$ where the distribution of ξ is independent of T.

For the PM algorithm, we have the following conditional CLT.

Theorem 2. *Let θ, ξ be fixed. Assume that $\vartheta \mapsto \varpi(y_1, u_{1,1}; \vartheta)$ and $\vartheta \mapsto \mathbb{E}\left(\varpi(Y_1, U_{1,1}; \vartheta)^9\right)$ are continuous at $\vartheta = \theta$ for any $(y_1, u_{1,1}) \in \mathsf{Y} \times \mathbb{R}^p$, $\vartheta \mapsto \gamma(\vartheta)$ is continuously differentiable at $\vartheta = \theta$ and $\mathbb{E}\left(\varpi(Y_1, U_{1,1}; \vartheta)^9\right) + \mathbb{E}\left(\gamma(Y_1; \theta)^4\right) < \infty$. For $N_T = \lceil \beta T^\alpha \rceil$ with $1/3 < \alpha \leq 1$, $\beta > 0$, $Y_t \overset{i.i.d.}{\sim} \mu$, $U \sim \overline{\pi}(\cdot \mid \theta)$ and $U' \sim m$ where U and U' are independent, we have*

$$T^{(\alpha-1)/2} R_T(\theta, \theta + \xi/\sqrt{T}) + T^{(1-\alpha)/2} \beta^{-1} \gamma(\theta)^2 \,\Big|\, \mathcal{Y}^T \Rightarrow \mathcal{N}\left(0, 2\beta^{-1}\gamma(\theta)^2\right). \tag{24}$$

This result shows that the loglikelihood ratio error in the PM case can only have a limiting variance of order 1 if N_T is proportional to T. The loglikelihood ratio estimator used by the CPM exhibits a markedly different behaviour if we consider the Crank-Nicolson proposal (8), $U' \sim K_{\rho_T}(U, \cdot)$, with

$$\rho_T = \exp\left(-\psi \frac{N_T}{T}\right), \tag{25}$$

for some $\psi > 0$. Let us denote by $\mathcal{F}^T$ the σ-field spanned by $\{Y_t; t \in 1:T\}$ and $\{U_{t,i}; t \in 1:T, i \in 1:N\}$. We also denote the Euclidean norm by $\|\cdot\|$ and write $\nabla_u f = (\partial_{u^1} f, ..., \partial_{u^p} f)'$ for a real-valued function $f : \mathbb{R}^p \to \mathbb{R}$ where $u = (u^1, ..., u^p)$.

Theorem 3. *Let θ, ξ be fixed. Let $Y_t \overset{i.i.d.}{\sim} \mu$, $U \sim \overline{\pi}(\cdot \mid \theta)$ and $U' \sim K_{\rho_T}(U, \cdot)$ where ρ_T is given by (25). Under Assumptions 1-6 in Section A.5 of the Supplementary Material, if $N_T \to \infty$ as $T \to \infty$ with $N_T/T \to 0$, we have*

$$R_T(\theta, \theta + \xi/\sqrt{T}) \,\Big|\, \mathcal{F}^T \Rightarrow \mathcal{N}\left(-\kappa(\theta)^2/2, \kappa(\theta)^2\right), \tag{26}$$

where

$$\kappa(\theta)^2 = 2\psi \mathbb{E}\left(\|\nabla_u \varpi(Y_1, U_{1,1}; \theta)\|^2\right). \tag{27}$$

Assumptions 1-6 are differentiability and integrability assumptions on $\varpi(y, u; \theta)$ with respect to y, u and θ. This result states that the limiting variance of the loglikelihood ratio for the CPM scheme at equilibrium is of order 1 when N_T grows sublinearly with T, although it will typically grow exponentially with p, the dimension of $U_{1,1}$. Moreover, the distribution of the loglikelihood ratio error is asymptotically independent of U, suggesting that the CPM chain is less prone to sticking than the PM chain at stationarity.

This conditional CLT has not been established for particle filters. For univariate state-space models, i.e., $k = 1$, we have observed experimentally on various stationary state-space models that a similar conditional CLT appears to hold. For multivariate state-space models, the CLT only appears to hold conditional upon $\mathcal{Y}^T$ when N_T grows at least at rate $T^{k/(k+1)}$; see Section 5.

4 Analysis and optimisation

4.1 Weak convergence in the large sample regime

The use of weak convergence techniques to analyse and optimise MCMC schemes was pioneered by Roberts et al. (1997) and has found numerous applications ever since; see, e.g., (Sherlock et al., 2015)

for a recent application to the PM method. The high-level idea behind this approach is to identify an appropriate asymptotic regime under which a component of the original Markov chain, rescaled appropriately, converges to a limiting process which is simpler to analyse and optimise. To the best of our knowledge, all previous contributions consider the asymptotic regime where $d \to \infty$, d being the parameter dimension, while T is fixed. In these scenarios, under time rescaling, the limiting Markov process is usually a diffusion. We analyse here the CPM scheme under the standard large sample regime of asymptotic statistics where d is fixed and $T \to \infty$. In this context, after space rescaling, the parameter component of the CPM chain, targeting the posterior $\pi_T(\theta)$ associated with the observations $Y_{1:T}$, converges towards a discrete-time Markov chain. Our analysis assumes the statistical model is regular enough to ensure that $\{\pi_T(\theta); T \geq 1\}$ can be approximated by normal densities which concentrate. Here $\pi_T(\theta)$ is interpreted as the density of a $\mathcal{Y}^T$-measurable random probability measure; see, e.g., (Berti et al., 2006; Crauel, 2003) for a formal definition. We write $\to_{\mathbb{P}^Y}$ to denote convergence in probability with respect to the law of $\{Y_t; t \geq 1\}$.

Assumption 1. *There exists a $d \times d$ positive definite matrix $\overline{\Sigma}$, a parameter value $\overline{\theta} \in \mathbb{R}^d$ and an $\mathbb{R}^d$-valued random sequence $\{\widehat{\theta}_T; T \geq 1\}$, $\widehat{\theta}_T$ being $\mathcal{Y}^T-$ measurable, such that as $T \to \infty$*

$$\int \left| \pi_T(\theta) - \varphi(\theta; \widehat{\theta}_T, \overline{\Sigma}/T) \right| \mathrm{d}\theta \to_{\mathbb{P}^Y} 0 \quad \textit{and} \quad \widehat{\theta}_T \to_{\mathbb{P}^Y} \overline{\theta}.$$

This assumption will be satisfied if a Berstein-von Mises theorem holds; see (van der Vaart, 2000, Section 10.2) for sufficient conditions.

Consider the stationary CPM chain $\{(\vartheta_n^T, \mathsf{U}_n^T); n \geq 0\}$ with proposal $q_T(\theta, \theta')$ targeting the random measure $\overline{\pi}_T(\mathrm{d}\theta, \mathrm{d}u) = \pi_T(\mathrm{d}\theta)\overline{\pi}_T(\mathrm{d}u|\theta)$ associated with the observations $Y_{1:T}$. By rescaling the parameter component of the CPM chain using $\widetilde{\vartheta}_n^T := \sqrt{T}(\vartheta_n^T - \widehat{\theta}_T)$, we obtain the stationary Markov chain $\{(\widetilde{\vartheta}_n^T, \mathsf{U}_n^T); n \geq 0\}$ with initial distribution $(\widetilde{\vartheta}_0^T, \mathsf{U}_0^T) \sim \widetilde{\pi}_T$ where

$$\widetilde{\pi}_T(\widetilde{\theta}, u) = \widetilde{\pi}_T(\widetilde{\theta})\widetilde{\pi}_T(u|\widetilde{\theta}), \quad \widetilde{\pi}_T(\widetilde{\theta}) = \pi_T(\widehat{\theta}_T + \widetilde{\theta}/\sqrt{T})/\sqrt{T}, \quad \widetilde{\pi}_T(u|\widetilde{\theta}) = \overline{\pi}_T(u|\widehat{\theta}_T + \widetilde{\theta}/\sqrt{T}), \tag{28}$$

and the associated proposal density for the parameter becomes

$$\widetilde{q}_T(\widetilde{\theta}, \widetilde{\theta}') = q_T(\widehat{\theta}_T + \widetilde{\theta}/\sqrt{T}, \widehat{\theta}_T + \widetilde{\theta}'/\sqrt{T})/\sqrt{T}. \tag{29}$$

We will assume here that we use a random walk proposal scaled appropriately.

Assumption 2. *The proposal density is of the form*

$$q_T(\theta, \theta') = \sqrt{T}\upsilon(\sqrt{T}(\theta' - \theta)), \tag{30}$$

where υ is a probability density on $\mathbb{R}^d$; that is $\theta' \sim q_T(\theta, \cdot)$ when $\theta' = \theta + \xi/\sqrt{T}$ with $\xi \sim \upsilon$.

Finally, we assume that a uniform version of the CLT of Theorem 3 holds in a neighbourhood of $\overline{\theta}$, where $\overline{\theta}$ is specified in Assumption 1. We denote by $d_{\mathrm{BL}}(\mu, \nu)$ the bounded Lipschitz metric between two probability measures μ and ν; see, e.g., (van der Vaart, 2000, p. 332) or Section A.9 of the Supplementary Material.

Assumption 3. *There exists a neighbourhood $N(\bar{\theta})$ of $\bar{\theta}$ such that the loglikelihood ratio error considered in Theorem 3 with $\xi \sim \upsilon(\cdot)$ satisfies as $T \to \infty$*

$$\sup_{\theta \in N(\bar{\theta})} \widetilde{\mathbb{E}} \left[d_{\mathrm{BL}} \left\{ \mathrm{Law}\left(R_T(\theta, \theta + \xi/\sqrt{T}) \middle| \mathcal{F}^T \right), \mathcal{N}\left(-\kappa(\theta)^2/2, \kappa(\theta)^2 \right) \right\} \middle| \mathcal{Y}^T \right] \to_{\mathbb{P}^Y} 0.$$

For the random effects model of Section 2.3.1, we prove that Assumption 3 holds under regularity conditions given in Section A.6 of the Supplementary Material.

Under Assumption 2, the proposal defined in (29) satisfies $\widetilde{q}_T(\widetilde{\theta}, \widetilde{\theta}') = \upsilon(\widetilde{\theta}' - \widetilde{\theta}) := \widetilde{q}(\widetilde{\theta}, \widetilde{\theta}')$. In this case, the corresponding transition kernel of the rescaled CPM chain is given by

$$Q_T\{(\widetilde{\theta}, u), (\mathrm{d}\widetilde{\theta}', \mathrm{d}u')\} = \widetilde{q}(\widetilde{\theta}, \mathrm{d}\widetilde{\theta}')K_{\rho_T}(u, \mathrm{d}u')\,\alpha_{Q_T}\{(\widetilde{\theta}, u), (\widetilde{\theta}', u')\} + \{1 - \varrho_{Q_T}(\widetilde{\theta}, u)\}\delta_{(\widetilde{\theta}, u)}(\mathrm{d}\widetilde{\theta}', \mathrm{d}u') \tag{31}$$

with acceptance probability

$$\alpha_{Q_T}\{(\widetilde{\theta},u),(\widetilde{\theta}',u')\}=\min\left\{1,\frac{\widetilde{\pi}_T(\widetilde{\theta}',u')\widetilde{q}(\widetilde{\theta}',\widetilde{\theta})K_{\rho_T}(u',u)}{\widetilde{\pi}_T(\widetilde{\theta},u)\widetilde{q}(\widetilde{\theta},\widetilde{\theta}')K_{\rho_T}(u,u')}\right\},$$

and corresponding rejection probability $1-\varrho_{Q_T}(\widetilde{\theta},u)$. The kernel Q_T is assumed to be $\mathcal{Y}^T$-measurable. Let $\Theta_T=\{\widetilde{\vartheta}_n^T;n\geq 0\}$ denote the non-Markov stationary space-rescaled parameter sequence arising from the CPM chain. The following result shows that the sequences $\{\Theta_T;T\geq 1\}$ converge weakly as $T\to\infty$ to a stationary Markov chain corresponding to the Penalty method – an "ideal" Monte Carlo technique which cannot be practically implemented (Ceperley et al., 1999; Nicholls et al., 2012).

Theorem 4. *If Assumptions 1, 2 and 3 hold and $\vartheta\mapsto\kappa(\vartheta)$ is locally Lipschitz at $\vartheta=\overline{\theta}$ then the random probability measures on $\left(\mathbb{R}^d\right)^\infty$ given by the laws of $\{\Theta_T;T\geq 1\}$ converge weakly in probability $\mathbb{P}^Y$ as $T\to\infty$ to the law of a stationary Markov chain $\{\widetilde{\vartheta}_n;n\geq 0\}$ defined by $\widetilde{\vartheta}_0\sim\mathcal{N}(0,\overline{\Sigma})$ and $\widetilde{\vartheta}_n\sim P(\widetilde{\vartheta}_{n-1},\cdot)$ for $n\geq 1$ with*

$$P(\widetilde{\theta},\mathrm{d}\widetilde{\theta}')=\widetilde{q}(\widetilde{\theta},\mathrm{d}\widetilde{\theta}')\alpha_P(\widetilde{\theta},\widetilde{\theta}')+\{1-\varrho_P(\widetilde{\theta})\}\delta_{\widetilde{\theta}}(\mathrm{d}\widetilde{\theta}'),\tag{32}$$

and

$$\alpha_P(\widetilde{\theta},\widetilde{\theta}')=\int\varphi\left(\mathrm{d}r;-\kappa^2/2,\kappa^2\right)\min\left\{1,\frac{\varphi(\widetilde{\theta}';0,\overline{\Sigma})\widetilde{q}(\widetilde{\theta}',\widetilde{\theta})}{\varphi(\widetilde{\theta};0,\overline{\Sigma})\widetilde{q}(\widetilde{\theta},\widetilde{\theta}')}\exp(r)\right\},$$

$1-\varrho_P(\widetilde{\theta})$ *being the corresponding rejection probability and* $\kappa:=\kappa(\overline{\theta})$.

The consequence of this result is that, as $T\to\infty$, only the asymptotic distribution of the loglikelihood ratio error at the central parameter value $\overline{\theta}$ impacts the acceptance probability of the limiting chain. For large T and a proposal of the form specified in Assumption 2, we thus expect some of the quantitative properties of the CPM kernel Q, where we now omit T from the notation, to be captured by the Markov kernel

$$\widehat{Q}(\theta,\mathrm{d}\theta')=q(\theta,\mathrm{d}\theta')\alpha_{\widehat{Q}}(\theta,\theta')+\{1-\varrho_{\widehat{Q}}(\theta)\}\delta_\theta(\mathrm{d}\theta'),\tag{33}$$

with

$$\alpha_{\widehat{Q}}(\theta,\theta')=\int\varphi(\mathrm{d}r;-\kappa^2/2,\kappa^2)\min\{1,r_{\text{MH}}(\theta,\theta')\exp(r)\},$$

where $1-\varrho_{\widehat{Q}}(\theta)$ is the corresponding rejection probability and r_{MH} is defined in (2). We have obtained (33) by using the change of variables $\theta=\widehat{\theta}_T+\widetilde{\theta}/\sqrt{T}$ and substituting the true target for its normal approximation in (32), hence removing a level of approximation.

4.2 A bounding Markov chain

We analyse here the stationary Markov chain with transition kernel $\widehat{Q}$ arising from our weak convergence analysis. To state our results, we need the following notation. For any real-valued measurable function h, probability measure μ and Markov kernel K on a measurable space $(E,\mathcal{E})$, we write $\mu(h)=\int_E h(x)\mu(\mathrm{d}x)$, $Kh(x)=\int_E K(x,\mathrm{d}x')h(x')$ and $K^nh(x)=\int_E\int_E K^{n-1}(x,\mathrm{d}z)K(z,\mathrm{d}x')h(x')$ for $n\geq 2$ with $K^1=K$. We also introduce the Hilbert space $L^2(\mu)=\left\{h:E\to\mathbb{R}\mid\mu\left(h^2\right)<\infty\right\}$ equipped with the inner product $\langle g,h\rangle_\mu=\int_E g(x)h(x)\mu(\mathrm{d}x)$. For any $h\in L^2(\mu)$, the autocorrelation at lag $n\geq 0$ is $\phi_n(h,K)=\left\langle\overline{h},K^n\overline{h}\right\rangle_\mu/\mu(\overline{h}^2)$ where $\overline{h}=h-\mu(h)$. The IACT associated with a function h under a Markov kernel K is given by $\mathrm{IF}(h,K)=1+2\sum_{n=1}^\infty\phi_n(h,K)$ and will be referred to subsequently as the *inefficiency*. For $\mu(\mathrm{d}x)=\mu(\mathrm{d}x_1,\mathrm{d}x_2)$, we will slightly abuse notation and write $\mathrm{IF}(h,K)$ instead of $\mathrm{IF}(g,K)$ when $g(x_1,x_2)=h(x_1)$ or $g(x_1,x_2)=h(x_2)$. When estimating $\mu(h)$, $n\mathrm{IF}(h,K)$ samples from a stationary Markov chain of μ-invariant transition kernel K are necessary to obtain an estimator of approximately the same precision as an average of n independent draws from μ; see, e.g., (Geyer, 1992).

We provide an upper bound on $\mathrm{IF}(h,\widehat{Q})$ which we exploit to provide guidelines on how to optimise the performance of the CPM scheme in Section 4.4. The inefficiency $\mathrm{IF}(h,\widehat{Q})$ is difficult to work with but we give an upper bound that only depends on $\mathrm{IF}(h,Q_{\text{MH}})$ and κ. To proceed, we introduce an auxiliary Markov kernel Q^* given by

$$Q^*(\theta,\mathrm{d}\theta')=\varrho_{\text{U}}(\kappa)Q_{\text{MH}}(\theta,\mathrm{d}\theta')+\{1-\varrho_{\text{U}}(\kappa)\}\delta_\theta(\mathrm{d}\theta'),\tag{34}$$

where Q_{MH} is defined in (1) and

$$\varrho_{\text{U}}(\kappa) = \int \varphi(\mathrm{d}r; -\kappa^2/2, \kappa^2) \min\{1, \exp(r)\} = 2\Phi(-\kappa/2). \tag{35}$$

We denote by $\bar{\varrho}_{\text{Q}^*}(\kappa)$, respectively $\bar{\varrho}_{\widehat{\text{Q}}}(\kappa)$, the average acceptance probability of Q^*, respectively $\widehat{Q}$, at stationarity. The kernel Q^* is a "lazy" version of Q_{MH} which satisfies the following properties.

Proposition 5. *The kernel Q^* is π-reversible and $\text{IF}(h, \widehat{Q}) \leq \text{IF}(h, Q^*)$ for any $h \in L^2(\pi)$, where*

$$\text{IF}(h, Q^*) = \{1 + \text{IF}(h, Q_{\text{MH}})\}/\varrho_{\text{U}}(\kappa) - 1, \tag{36}$$

with equality when $\varrho_{\text{MH}}(\theta) = 1$ for all $\theta \in \Theta$, and

$$\bar{\varrho}_{\text{Q}^*}(\kappa) = \varrho_{\text{U}}(\kappa)\,\pi(\varrho_{\text{MH}}) \leq \bar{\varrho}_{\widehat{\text{Q}}}(\kappa). \tag{37}$$

Moreover, Q^* is geometrically ergodic if Q_{MH} is geometrically ergodic.

For any π or $\overline{\pi}$-invariant Markov kernel K, we define the relative inefficiency $\text{RIF}(h, K)$ and the auxiliary relative computing time $\text{ARCT}(h, K)$ with respect to the MH kernel Q_{MH} using the exact likelihood by

$$\text{RIF}(h, K) := \frac{\text{IF}(h, K)}{\text{IF}(h, Q_{\text{MH}})}, \qquad \text{ARCT}(h, K) := \sqrt{\frac{\text{RIF}(h, K)}{\kappa^2 \varrho_{\text{U}}(\kappa)}}. \tag{38}$$

We next minimise $\text{ARCT}(h, Q^*)$, an upper bound on $\text{ARCT}(h, \widehat{Q})$, with respect to κ – this quantity is a component of the function we need to minimise in order to optimise the performance of the CPM algorithm; see Section 4.4.

Proposition 6. *The following results hold:*

1. *If $\text{IF}(h, Q_{\text{MH}}) = 1$, then*
$$\text{RIF}(h, Q^*) = \{2 - \varrho_{\text{U}}(\kappa)\}/\varrho_{\text{U}}(\kappa),$$
and $\text{ARCT}(h, Q^*)$ is minimised at $\kappa = 1.35$, at which point $\varrho_{\text{U}}(\kappa) = 0.50$, $\text{RIF}(h, Q^*) = 2.99$ and $\text{ARCT}(h, Q^*) = 1.81$.

2. *As $\text{IF}(h, Q_{\text{MH}}) \longrightarrow \infty$,*
$$\text{RIF}(h, Q^*) = 1/\varrho_{\text{U}}(\kappa),$$
and $\text{ARCT}(h, Q^*)$ is minimised at $\kappa = 1.50$, at which point $\varrho_{\text{U}}(\kappa) = 0.43$, $\text{RIF}(h, Q^*) = 2.20$ and $\text{ARCT}(h, Q^*) = 1.47$.

3. *$\text{RIF}(h, Q^*)$ and $\text{ARCT}(h, Q^*)$ are decreasing functions of $\text{IF}(h, Q_{\text{MH}})$. The minimising argument rises monotonically from 1.35 to 1.50 as $\text{IF}(h, Q_{\text{MH}})$ increases from 1 to ∞.*

Figure 1 displays $\varrho_{\text{U}}(\kappa)$, $\text{RIF}(h, Q^*)$ and $\text{ARCT}(h, Q^*)$ against κ. The two scenarios displayed are for $\text{IF}(h, Q_{\text{MH}}) = 1$, corresponding to the "perfect" proposal case where $q(\theta, \theta') = \pi(\theta')$, and for the limiting case where $\text{IF}(h, Q_{\text{MH}}) \longrightarrow \infty$. These correspond to parts 1 and 2 of Proposition 6. From Figure 1, it is also clear that $\text{ARCT}(h, Q^*)$, for both scenarios, is fairly flat as a function of κ. The function only approximately doubles relative to the minimum at $\kappa = 1$ or 4.

4.3 A lower bound on the integrated autocorrelation time

We stress here that Theorem 4 does not imply that the inefficiency of the CPM scheme converges, as $T \to \infty$, to the inefficiency of the limiting chain identified therein. In fact, whereas Theorem 4 holds whenever $N_T \to \infty$ and $N_T = o(T)$ as $T \to \infty$, our next result suggests that N_T must grow at least as fast as $\sqrt{T}$ for the inefficiency of the CPM scheme to remain bounded. To simplify the presentation in this section, we assume further on that $d = 1$.

In the CPM context, the sequence of auxiliary variables $\{\mathsf{U}_n; n \geq 0\}$ evolve at a much slower scale than $\{\vartheta_n; n \geq 0\}$ as it is driven by the proposal K_{ρ_T}, where ρ_T is given by (25). When N_T grows too slowly

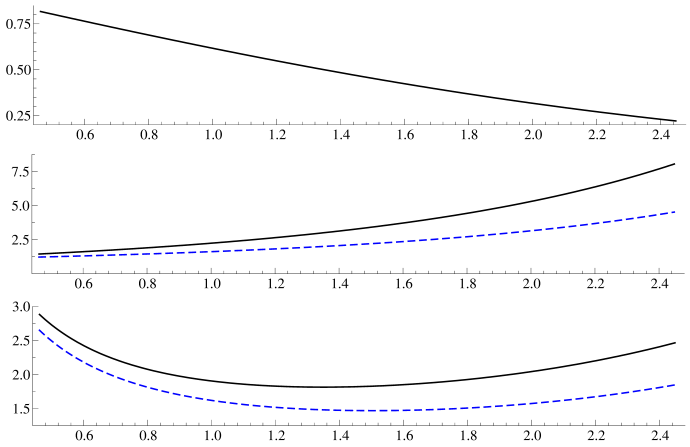

Figure 1: Illustrations of Proposition 6. Top: Acceptance probability $\varrho_{\mathrm{U}}(\kappa)$ against κ. Middle: Relative inefficiency $\mathrm{RIF}\,(h, Q^*)$ against κ (solid line $\mathrm{IF}(h, Q_{\mathrm{MH}}) = 1$, dashed line $\mathrm{IF}(h, Q_{\mathrm{MH}}) \to \infty$). Bottom: Auxiliary relative computing time $\mathrm{ARCT}\,(h, Q^*)$ against κ (solid line $\mathrm{IF}(h, Q_{\mathrm{MH}}) = 1$, dashed line $\mathrm{IF}(h, Q_{\mathrm{MH}}) \to \infty$).

with T, we expect and observe empirically that the inefficiency $\mathrm{IF}(h, Q_T)$, for any function h, is of the same order as the inefficiency of $\{\mathbb{E}\,[h(\vartheta_n)|\mathsf{U}_n]\,; n \geq 0\}$. Moreover, under regularity conditions, see e.g. (Doucet et al., 2013, Lemma 2), we have for large T

$$\mathbb{E}\,[h(\vartheta_n)|\mathsf{U}_n] = h(\widehat{\theta}_T) + \frac{\overline{\Sigma}}{2T}\nabla_{\vartheta,\vartheta} h(\widehat{\theta}_T) + \frac{\overline{\Sigma}}{T}\nabla_{\vartheta} h(\widehat{\theta}_T)\ \Psi(\widehat{\theta}_T, \mathsf{U}_n) + O_{\mathbb{P}}\left(T^{-2}\right), \tag{39}$$

where

$$\Psi(\widehat{\theta}_T, U) = \nabla_{\vartheta} \log\{\widehat{p}(Y_{1:T} \mid \widehat{\theta}_T, U)/p(Y_{1:T} \mid \widehat{\theta}_T)\} \tag{40}$$

is the error in the simulated score at $\widehat{\theta}_T$, and will be referred to as the score error. As a first step, we obtain a lower bound on $\mathrm{IF}(\Psi, Q_T)$.

Proposition 7. *Under regularity conditions given in Section A.10 of the Supplementary Material, there exists a constant $C > 0$ such that* $\mathrm{IF}(\Psi, Q_T) \geq C\mathbb{V}_{\overline{\pi}_T}(\Psi)\ \ \mathbb{P}^Y - \text{a.s.}$

It follows from calculations similar to Section A.11 in the Supplementary Material, see also (Lindsten and Doucet, 2016, Proposition 3), that under regularity conditions there exists $A > 0$ such that $\mathbb{V}_{\overline{\pi}_T}(\Psi) \sim AT/N\ \mathbb{P}^Y - \text{a.s.}$ By combining (39) and Proposition 7, we thus expect the inefficiency of $\{\mathbb{E}\,[h(\vartheta_n)|\mathsf{U}_n]\,; n \geq 0\}$ to be lower bounded by a term of order

$$\frac{\mathrm{IF}(\Psi, Q_T)\ \mathbb{V}_{\overline{\pi}_T}(\Psi/T)}{\mathbb{V}_{\overline{\pi}_T}(h)} \geq B\frac{T}{N_T}\frac{T^{1-\alpha}}{T^2}T = BT^{1-2\alpha}$$

for $N_T = \lceil \beta T^{\alpha} \rceil$, some constant $B > 0$ and T large enough. This result suggests that a necessary condition for $\mathrm{IF}(h, Q_T)$ to remain finite as $T \to \infty$ is to have N_T growing at least at rate $\sqrt{T}$. This is validated by the experimental results of Section 5 which also suggest that this rate is sufficient.

4.4 Optimization

We provide a heuristic to select the parameters of the CPM scheme so as to optimise its performance which is validated by experimental results in Section 5. Again, we set $d = 1$ for simplicity's sake. For a test function $h : \Theta \to \mathbb{R}$, we want to minimise

$$\mathrm{CT}(h, Q_T) = N_T \times \mathrm{IF}(h, Q_T), \tag{41}$$

where the factor N_T arises from the fact that the computational cost of the likelihood estimator is proportional to N_T for random effects models. The results of Section 4.3 suggest that we should choose

the number of Monte Carlo samples to scale as $N_T = \beta T^{1/2}$ so that $\rho_T = \exp\left(-\psi\beta T^{-1/2}\right)$. It remains to determine ψ and β.

To evaluate (41), we first decompose the functional of interest evaluated at the parameter at the n-th iteration as

$$h(\vartheta_n) = f(\mathsf{U}_n) + g(\vartheta_n, \mathsf{U}_n),$$

where

$$f(\mathsf{U}) := \mathbb{E}_{\bar{\pi}_T}\left[h(\vartheta)|\mathsf{U}\right], \qquad g(\vartheta, \mathsf{U}) := h(\vartheta) - \mathbb{E}_{\bar{\pi}_T}\left[h(\vartheta)|\mathsf{U}\right]. \tag{42}$$

It is easy to check that

$$\mathbb{V}_{\bar{\pi}_T}(h)\,\mathrm{IF}(h, Q_T) \leq 2\mathbb{V}_{\bar{\pi}_T}(f)\mathrm{IF}(f, Q_T) + 2\mathbb{V}_{\bar{\pi}_T}(g)\mathrm{IF}(g, Q_T).$$

Assumption 1 combined with mild regularity assumptions on h and integrability conditions shows that $\mathbb{V}_{\bar{\pi}_T}\left(h\left(\vartheta_n\right)\right) \approx \overline{\Sigma}_h/T$, where $\overline{\Sigma}_h = |h'(\bar{\theta})|^2\overline{\Sigma}$. Since $f(\mathsf{U}_n)$ and $g(\vartheta_n, \mathsf{U}_n)$ are clearly uncorrelated, it follows that $\mathbb{V}_{\bar{\pi}_T}(h) = \mathbb{V}_{\bar{\pi}_T}(f) + \mathbb{V}_{\bar{\pi}_T}(g)$. From (39) we have $\mathbb{V}_{\bar{\pi}_T}(f) \approx \overline{\Sigma}^2\mathbb{V}_{\bar{\pi}_T}(\Psi/T) \approx \overline{\Sigma}_f/(TN_T)$, therefore

$$\mathbb{V}_{\bar{\pi}_T}(g) \approx \frac{\overline{\Sigma}_h}{T} - \frac{\overline{\Sigma}_f}{TN_T} \approx \frac{\overline{\Sigma}_h}{T}.$$

Using the reasoning of Section 4.3 and the calculations above we obtain

$$\begin{aligned}\mathrm{IF}\left(h, Q_T\right) &\leq \frac{2}{\overline{\Sigma}_h}\left(\mathbb{V}_{\bar{\pi}_T}\left(\sqrt{T}f\right)\mathrm{IF}\left(f, Q_T\right) + \mathbb{V}_{\bar{\pi}_T}\left(\sqrt{T}g\right)\mathrm{IF}\left(g, Q_T\right)\right) \\ &\approx \frac{2}{\overline{\Sigma}_h}\left(\frac{\overline{\Sigma}_f}{N_T}\mathrm{IF}\left(\Psi, Q_T\right) + \overline{\Sigma}_h\mathrm{IF}\left(g, Q_T\right)\right).\end{aligned} \tag{43}$$

Proposition 7 states that $\mathrm{IF}(\Psi, Q_T)$ is of order at least T/N_T in probability as $T \to \infty$. Numerical results suggest that in fact we have $\mathrm{IF}(\Psi, Q_T) \approx A/\left(\delta_T\varrho_{\mathrm{U}}(\kappa)\right)$ where $\delta_T = \psi N_T/T = -\log\rho_T$ as illustrated in Section 5.1, Figure 5. Hence, by substituting this expression of $\mathrm{IF}(\Psi, Q_T)$ in (43), it follows that

$$\mathrm{IF}\left(h, Q_T\right) \lesssim \frac{2}{\overline{\Sigma}_h}\left(\frac{\overline{\Sigma}_f}{N_T}\frac{A}{\delta_T\varrho_{\mathrm{U}}(\kappa)} + \overline{\Sigma}_h\,\mathrm{IF}\left(g, Q_T\right)\right),$$

where the symbol $\lesssim$ means that an approximation has been used. It can also be observed empirically from Figure 4, described in Section 5.1, that the autocorrelations of $g(\vartheta_n, \mathsf{U}_n)$ decay exponentially, at a rate independent of T. We expect that, at least approximately, we have $\mathrm{IF}\left(g, Q_T\right) \approx \mathrm{IF}(h, \widehat{Q}_T)$ in probability. Therefore overall, for some constant $B > 0$, we have that

$$\mathrm{IF}\left(h, Q_T\right) \lesssim 2\left(\frac{B}{\varrho_{\mathrm{U}}(\kappa)\delta_T N_T} + \mathrm{IF}\left(h, \widehat{Q}_T\right)\right). \tag{44}$$

We are interested in optimizing $\mathrm{CT}(h, Q_T) = N_T \times \mathrm{IF}(h, Q_T)$ with respect to ψ and β where we recall from (27) that $\delta_T = \psi N_T/T = \psi\beta/\sqrt{T} = (\kappa^2\beta)/(\gamma^2\sqrt{T})$ as $\kappa^2 = \psi\gamma^2$. Therefore

$$\mathrm{CT}(h, Q_T) \lesssim 2T^{1/2}\left(\frac{C}{\beta\varrho_{\mathrm{U}}(\kappa)\kappa^2} + \beta\mathrm{IF}\left(h, \widehat{Q}_T\right)\right), \tag{45}$$

where $C = B\gamma^2$, and the upper bound on $\mathrm{CT}(h, Q_T)$ is minimised at

$$\beta^* = \sqrt{\frac{C}{\varrho_{\mathrm{U}}(\kappa)\kappa^2\mathrm{IF}\left(h, \widehat{Q}_T\right)}}.$$

By plugging β^* in the right hand side of (45), we obtain by Proposition 5

$$\mathrm{CT}(h, Q_T) \lesssim 4\sqrt{C\,\mathrm{IF}(h, Q_{\mathrm{MH}})\,T} \times \mathrm{ARCT}\left(h, \widehat{Q}_T\right) \lesssim 4\sqrt{C\,\mathrm{IF}(h, Q_{\mathrm{MH}})\,T} \times \mathrm{ARCT}(h, Q_T^*) \tag{46}$$

where ARCT was introduced in (38). In practice we minimise $\mathrm{ARCT}(h, Q_T^*)$ with respect to κ, following Proposition 6. The minimiser $\hat{\kappa}$ is a function of $\mathrm{IF}(h, Q_{\mathrm{MH}})$ which varies only slightly as $\mathrm{IF}(h, Q_{\mathrm{MH}})$ varies from 1 to ∞ as observed in Figure 1. Consequently, we propose the following procedure to optimise

the performance of CPM. Let T be fixed and large enough for the asymptotic assumptions to hold approximately. First, we choose a candidate value for N and determine $\hat{\psi}$ such that the standard deviation of the log-likelihood ratio estimator around the mode of the posterior, estimated through a preliminary run, satisfies $\hat{\kappa} \approx 1.4$. Second, fixing ψ to $\hat{\psi}$, we evaluate for several values of β the computation time $\mathrm{CT}(h, Q_T)$ which we assume is of the form of the upper bound (45), i.e.,

$$\mathrm{CT}(h, Q_T) = C_0/\beta + C_1\beta, \tag{47}$$

with κ and T kept constant; see Figure 6 in Section 5.1 for empirical results. This function is minimised for $\beta = \sqrt{C_0/C_1}$. Practically we only evaluate $\mathrm{CT}(h, Q_T)$ on a subset of the data. We then estimate through regression the constants C_0 and C_1 by $\hat{C}_0$ and $\hat{C}_1$ which in turn provide the following estimate of β

$$\hat{\beta} = \sqrt{\hat{C}_0/\hat{C}_1}. \tag{48}$$

We examine in Section 5.1 the assumptions made here, illustrate this procedure and demonstrate its robustness.

5 Applications

5.1 Random effects model

We illustrate the performance of the PM and CPM schemes on a simple Gaussian random effects model where

$$X_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, 1), \qquad Y_t | X_t \sim \mathcal{N}(X_t, 1). \tag{49}$$

We are interested in estimating θ (which has a true value of 0.5) to which we assign a zero-mean Gaussian prior with large variance. In this scenario, the likelihood is known as $Y_t \sim \mathcal{N}(\theta, 2)$. This allows for detailed experimental analysis of the loglikelihood error and the loglikelihood ratio error. This also allows us to implement the MH algorithm with the true likelihood. The same normal random walk proposal is used for all three schemes (MH, PM and CPM) and the following unbiased estimator of the likelihood is used for the PM and CPM schemes:

$$\widehat{p}(y_{1:T} \mid \theta, U) = \prod_{t=1}^{T} \widehat{p}(y_t \mid \theta, U_t), \qquad \widehat{p}(y_t \mid \theta, U_t) = \frac{1}{N}\sum_{i=1}^{N} \varphi\left(y_t; \theta + U_{t,i}, 1\right), \qquad U_{t,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \tag{50}$$

The inefficiency is estimated for all three schemes for $h(\theta) = \theta$ using $1 + 2\sum_{n=1}^{L} \widehat{\phi}_n$ where $\widehat{\phi}_n$ is the estimated correlation for θ at lag n and L is a suitable cutoff value. We use the notation $Z = \log\{\widehat{p}(y_{1:T} \mid \theta, U)/p(y_{1:T} \mid \theta)\}$ and $W = \log\{\widehat{p}(y_{1:T} \mid \theta', U')/p(y_{1:T} \mid \theta')\}$ where $\theta' \sim q(\theta, \cdot)$, $U' \sim K_\rho(U, \cdot)$ and write $R = W - Z$ for $R_T(\theta, \theta')$ defined in (23).

As discussed in Section 4, for large datasets, the relative inefficiency $\mathrm{RIF} = \mathrm{IF}/\mathrm{IF}_{\mathrm{MH}}$ and associated relative computing time $\mathrm{RCT} = N \times \mathrm{RIF}$ of the CPM scheme depend on the standard deviation κ of R at stationarity and the correlation parameter ρ. To validate experimentally the results of Section 3, we first analyse the case where $T = 8192$ in more detail. We run CPM using a random walk proposal for $N = 80$ and $\rho = 0.9963$, so that $\kappa = 1.145$. The draws of W and Z at equilibrium, together with R, are displayed in Figure 2. The draws of Z are approximately distributed according to $\mathcal{N}(\sigma^2/2, \sigma^2)$ (middle top), where the variance σ^2 is high. The draws of R appear uncorrelated (in unreported tests) and their histogram is indistinguishable from the expected theoretical distribution $\mathcal{N}(-\kappa^2/2, \kappa^2)$ established in Theorem 3 (middle bottom). This is in agreement with Theorem 1, equation (22), the posterior of θ being concentrated. The resulting draws and correlogram (bottom left and right) of θ demonstrate low persistence.

For the PM scheme, it is necessary to take $N = 5000$ samples to ensure that the variance of Z evaluated at a central value $\widehat{\theta}$ is approximately one (Doucet et al., 2015). We next validate experimentally the theoretical results of Section 4 by investigating the performance of CPM for this dataset, varying N, and thus also $\kappa^2 = \mathbb{V}(R)$, while keeping $\rho = 0.9963$. Figure 3 displays the values of RIF and RCT against κ as well as the marginal acceptance probabilities, showing that RCT is approximately minimised around $\kappa = 1.6$ close to the minimising argument of $\mathrm{ARCT}(h, Q_T^*)$ established in Proposition 6 which satisfies

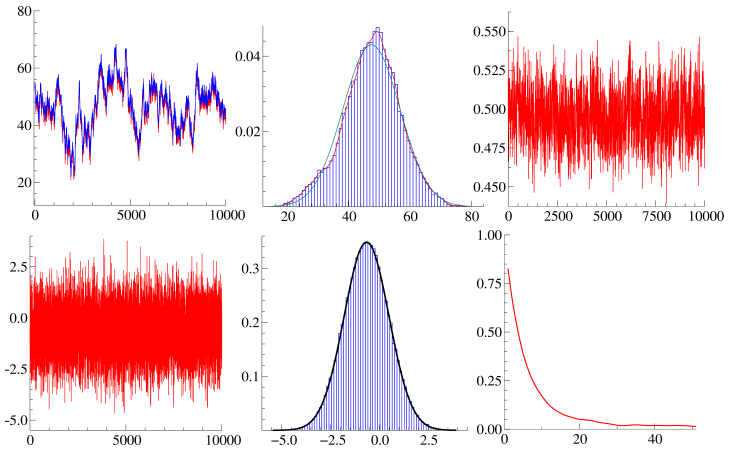

Figure 2: Random effects model using CPM: $T = 8192$, $N = 80$, $\rho = 0.9963$. Left: the first 10000 iterations of W (blue) and Z (red) (top), the difference R (bottom). Middle: Histograms of Z (top) and R (bottom) and the theoretical Gaussian densities. Right: draws of θ (top) and the corresponding correlogram (bottom).

(46). The bottom two plots show that $\log(\kappa^2)$ decreases linearly with $\log(N)$ as expected (bottom right) and that the marginal probability of acceptance in the CPM scheme is close to the asymptotic lower bound (bottom left) given by (37). From these experimental results, it is clear that for all values of N considered, the gains of the CPM scheme over the PM method in terms of RCT are very significant. The optimal value of N for the CPM scheme is 35 ($\kappa = 1.6$) which gives RCT = 61 against a value of RCT = 14100 for the PM scheme. Consequently, the PM method would take more than 200 times as long in computation time to produce an estimate of the posterior mean of θ of the same accuracy.

We next investigate the performance of the CPM method when T and $N = \beta\sqrt{T}$ vary while ψ, equivalently ρ, is scaled such that κ is approximately constant. The results are recorded in Table 1. They suggest that the scaling $N = \beta\sqrt{T}$ is successful as $\mathrm{IF}_{\mathrm{CPM}}$ appears to stabilize whereas the scaling $N = \beta T$ is necessary for $\mathrm{IF}_{\mathrm{PM}}$ to stabilize. Experimental results not reported here confirm that if N grows at a slower rate than $\sqrt{T}$, then $\mathrm{IF}_{\mathrm{CPM}}$ increases without bound with T.

T	N	ρ	κ^2	$\bar{\varrho}_{\mathrm{MH}}$	$\mathrm{IF}_{\mathrm{MH}}$	$\bar{\varrho}_{\mathrm{CPM}}$	$\mathrm{IF}_{\mathrm{CPM}}$	$\mathrm{RIF}_{\mathrm{CPM}}$
1024	19	0.9894	2.0	0.71	10.71	0.48	43.26	4.04
2048	28	0.9925	1.9	0.69	8.21	0.49	38.50	4.61
4096	39	0.9947	1.7	0.72	11.75	0.51	21.01	1.79
8192	56	0.9962	1.8	0.81	15.61	0.50	24.25	1.55
16384	79	0.9974	1.8	0.70	9.37	0.50	20.05	2.14

Table 1: Random effects model. Inefficiency and acceptance probabilities for MH and CPM, $N = \beta\sqrt{T}$ and ρ selected such that κ^2 is approximately constant.

We now justify empirically some of the assumptions made in Section 4 to guide the selection of the parameters ψ and β. First, we show that the CPM process can be thought of as a combination of two different processes: a 'slow' moving component $f(\mathsf{U}_n) \approx \hat{f}(\mathsf{U}_n) = \widehat{\theta}_T + \overline{\Sigma} T^{-1} \Psi(\widehat{\theta}_T, \mathsf{U}_n)$, the modified score error associated to the score error $\Psi(\widehat{\theta}_T, \mathsf{U}_n)$ defined in (40), and a 'fast' component $g(\vartheta_n, \mathsf{U}_n) = \vartheta_n - f(\mathsf{U}_n) \approx \hat{g}(\vartheta_n, \mathsf{U}_n) = \vartheta_n - \hat{f}(\mathsf{U}_n)$. We display these components for a CPM run and the associated correlograms in Figure 4 for fixed κ. We also illustrate in Figure 5 that $\mathrm{IF}(\Psi, Q_T) \approx A/\left(\delta_T \varrho_{\mathrm{U}}(\kappa)\right)$ where $\delta_T = \psi N_T / T = -\log \rho_T$. The optimisation scheme developed in Section 4.4 essentially selects β such that the asymptotic variances of both the slow and fast components $\hat{f}(\mathsf{U}_n)$ and $\hat{g}(\vartheta_n, \mathsf{U}_n)$ are of the same order.

To apply the optimization procedure, we first run the algorithm for $N = 20$ and tune ψ to get $\hat{\kappa} \approx 1.4$. For the resulting value $\hat{\psi}$, we then evaluate $\mathrm{CT}_{\mathrm{CPM}} = N \times \mathrm{IF}_{\mathrm{CPM}}$ for various values of β and perform a regression based on (47)-(48). Practically, we only use a subset of the data to perform this optimisation

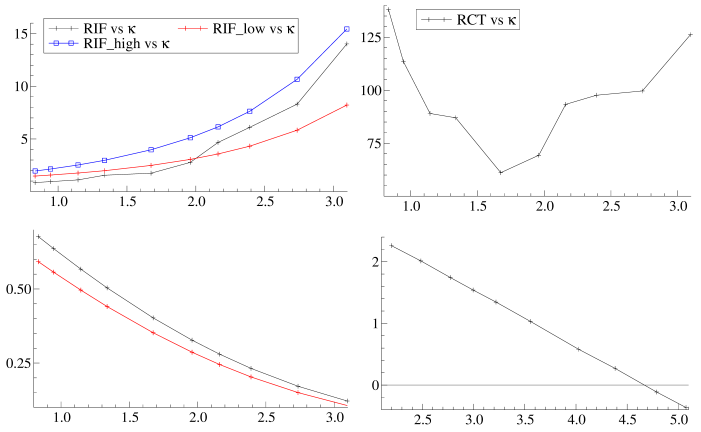

Figure 3: Random effects model using CPM: $T = 8192$, ρ fixed and various N. RIF_{CPM} (top left) and RIF_{Q^*} for $\text{IF}(h, Q_{\text{MH}}) = 1$ and $\text{IF}(h, Q_{\text{MH}}) = \infty$ against κ, see Proposition 6. RCT_{CPM} against κ (top right). The acceptance probability of the CPM and the theoretical lower bound, of (37), against κ (bottom left). $\log(\kappa^2)$ against $\log(N)$ (bottom right).

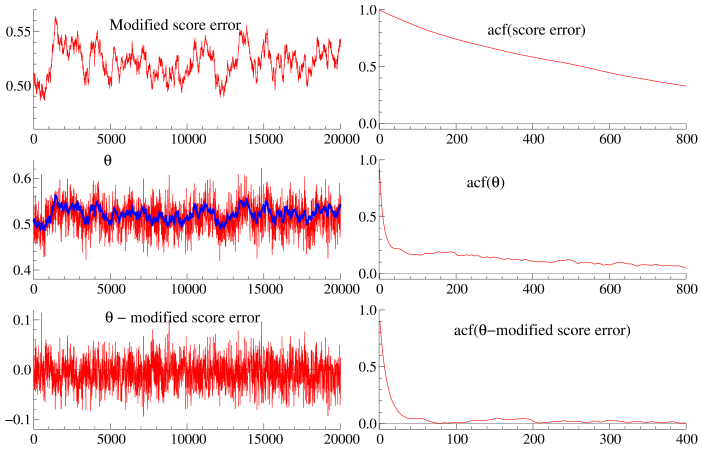

Figure 4: Random effects model using CPM: $T = 2560$, $\beta = 0.12$, $N = 6$, $\rho = 0.9977$. Top: modified score error $\hat{f}(\mathsf{U}_n)$ (left) and its correlogram (right). Middle: parameter ϑ_n (red) and modified score error (blue) (left) and correlogram ϑ_n (right). Bottom: residual $\hat{g}(\vartheta_n, \mathsf{U}_n)$ (left) and correlogram (right).

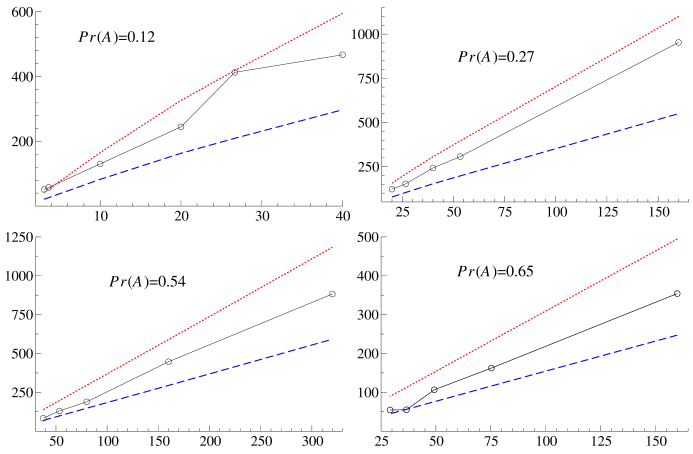

Figure 5: Random effects model using CPM: $T = 320$. Inefficiency of the score error (black line) plotted against $1/\delta$ for four different values of $\kappa^2 = 9.5, 4.9, 1.42, 0.75$ from top left to bottom right clockwise and corresponding acceptance probability $\bar{\varrho}_{\text{CPM}}$. Upper bound $2/(\delta\bar{\varrho}_{\text{CPM}})$ (dotted red) and lower bound $1/(\delta\bar{\varrho}_{\text{CPM}})$ (dotted blue).

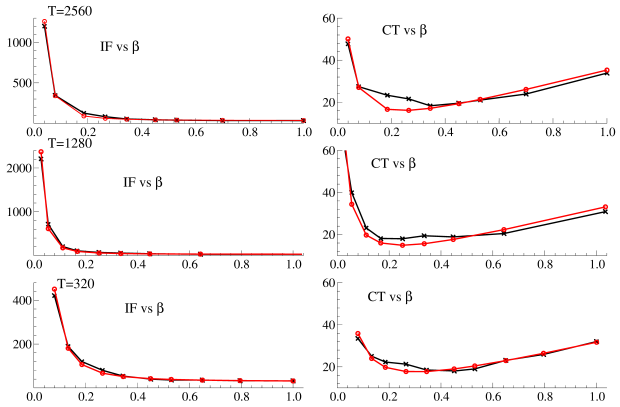

Figure 6: Random effects model. IF and CT as a function of β. Top to bottom: $T = 2560, 1280, 320$. Left: IF = IACT vs β. Right: CT = IF $\times\, \beta$ vs β. The regression fit based upon estimated CT is included in red.

to speed up computation. The results are fairly insensitive to the size of this subset as illustrated in Figure 6 and suggest selecting β around 0.25.

5.2 Heston stochastic volatility model

We investigate here the empirical performance of CPM on the Heston model (Heston, 1993; Chopin and Gerber, 2017), a popular stochastic volatility model with leverage which is a partially observed diffusion model. The logarithm of the observed price $P(t)$ evolves according to

$$\begin{aligned} \mathrm{d}\log P(t) &= \sigma(t)\mathrm{d}B(t), \\ \mathrm{d}\sigma^2(t) &= \upsilon\left\{\mu - \sigma^2(t)\right\}\mathrm{d}t + \omega\sigma(t)\,\mathrm{d}W(t), \end{aligned}$$

where $\sigma(t)$ is a stationary latent spot stochastic volatility process such that $\sigma^2(t) \sim \mathcal{G}(\alpha, \beta)$ where $\mathcal{G}(\alpha, \beta)$ is the gamma distribution of shape $\alpha = 2\mu\upsilon/\omega^2$ and rate $\beta = 2\upsilon/\omega^2$. The Brownian motions $B(t)$ and $W(t)$ are correlated with $\chi = \text{corr}\{B(t), W(t)\}$. The returns $Y_s = \log P(\tau_s) - \log P(\tau_{s-1})$ are observed at equally spaced times $\tau_0 < \cdots < \tau_T$, where $\triangle = \tau_s - \tau_{s-1}$ for all $s = 1, ..., T$. Condition upon

the volatility $\sigma^2(t)$ and driving process $W(t)$, we have

$$Y_s \sim \mathcal{N}\left\{\chi\gamma_s; (1-\chi^2)\sigma_s^{2*}\right\}, \tag{51}$$

$$\sigma_s^{2*} = \int_{\tau_{s-1}}^{\tau_s} \sigma^2(t)\mathrm{d}t, \quad \gamma_s = \int_{\tau_{s-1}}^{\tau_s} \sigma(t)\mathrm{d}W(t). \tag{52}$$

To perform inference, we first reparameterise the model in terms of $x(t) = \log\sigma^2(t)$. We apply Itô's lemma to $x(t)$ and discretise the resulting diffusion using an Euler scheme. We write $x_i^s = x\left(\tau_s + \epsilon i\right)$, where $\epsilon = \triangle/I$ for $i = 0, ..., I$ so that $x_I^s = x_0^{s+1}$. The evolution of these latent variables is given by

$$x_{i+1}^s = x_i^s + \epsilon\left[\upsilon\left\{\mu e^{-x_i^s} - 1\right\} - \frac{\omega^2}{2}e^{-x_i^s}\right] + \sqrt{\epsilon}\omega e^{-x_i^s/2}\eta_i,$$

where $\eta_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ for $i = 0, ..., I-1$. Under the Euler scheme, the returns satisfy

$$Y_s \sim \mathcal{N}\left\{\chi\widehat{\gamma}_s; (1-\chi^2)\widehat{\sigma}_s^{2*}\right\}, \tag{53}$$

$$\widehat{\sigma}_s^{2*} = \epsilon\textstyle\sum_{i=1}^I \exp(x_i^s), \quad \widehat{\gamma}_s = \sqrt{\epsilon}\textstyle\sum_{i=1}^I \exp(x_i^s/2)\eta_i, \tag{54}$$

where $\widehat{\sigma}_s^{2*}$ and $\widehat{\gamma}_s$ are the Euler approximations of the expressions in (52). We are interested in inferring $\theta = (\mu, \upsilon, \omega, \chi)$ given $T = 4000$ daily returns $y_{1:T}$ from the S&P 500 index from 15/08/1990 to 03/07/2006. We use here $I = 10$. Although the state is scalar, it is very difficult to perform inference using standard MCMC techniques as this involves $T \times I = 40000$ highly correlated latent variates.

We first run the CPM scheme keeping the parameter fixed at the posterior mean $\hat{\theta}$, estimated from a full CPM run, and only updating the auxiliary variables. We display the histograms of $Z = \log\widehat{p}(y_{1:T} \mid \hat{\theta}, U)$, $W = \log\widehat{p}(y_{1:T} \mid \hat{\theta}, U')$ and $R = \log\{\widehat{p}(y_{1:T} \mid \hat{\theta}, U')/\widehat{p}(y_{1:T} \mid \hat{\theta}, U)\}$ in Figure 7 for $N = 80$ and $N = 300$ using the parameters given in Table 2. We observe that R is approximately distributed according to $\mathcal{N}(-\kappa^2/2, \kappa^2)$ for $\kappa = 1.35$ in both cases. Additionally the sequence of estimates is almost uncorrelated across CPM iterations.

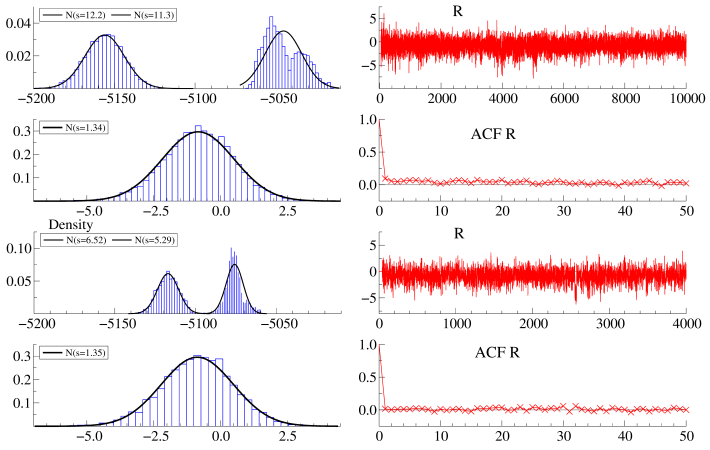

Figure 7: Histograms of W, Z for $N = 80$ (1st left), $N = 300$ (3rd left), histograms of R for $N = 80$ (2nd left), $N = 300$ (4th left). R across CPM iterations and associated correlograms for $N = 80$ (1st right, 2nd right), $N = 300$ (3rd right, 4th right).

Using $N = 300$, we first select $\psi = 0.125$ to achieve $\kappa = 1.4$ at $\hat{\theta}$. We then run the CPM scheme using a random walk proposal for other values of N, $N = \beta\sqrt{T}$, and compute CT $= N \times$ IF. These results are summarized in Table 2. The posterior estimates are in very close agreement across the different values of N. In unreported results, we observe empirically that the dependence of CT on β for parameters $(\mu, \phi := e^{-\upsilon}, \omega, \chi)$ matches (47) which can be optimized, suggesting that an optimal value of N around 70-80. As in the random effects scenario, we observe on datasets of increasing length that the scaling $N = \beta\sqrt{T}$ is successful as IF_{CPM} appears to stabilize. In this context, the PM method is extremely expensive computationally as we need approximately $N = 20000$ to obtain a standard deviation of Z around one (Doucet et al., 2015), our implementation taking 7 minutes per iteration to run on a standard desktop. In terms of CT, the CPM scheme is approximately 100 times more efficient than the PM scheme.

$\mathbb{E}(\theta)$ (SD(θ))	μ	ϕ	ω	χ	CPM ρ
$N = 80$	1.258 (0.098)	0.981 (0.0027)	0.142 (0.0099)	-0.676 (0.027)	0.9975
$N = 150$	1.253 (0.098)	0.981 (0.0028)	0.142 (0.0105)	-0.672 (0.034)	0.9953
$N = 300$	1.255 (0.099)	0.981 (0.0028)	0.142 (0.0110)	-0.671 (0.032)	0.9907

CT(θ)	μ	ϕ	ω	χ	$\bar{\varrho}_{\text{CPM}}$
$N = 80$	9995	12555	13571	33794	0.276
$N = 150$	19691	20256	17931	32588	0.272
$N = 300$	32970	30432	35103	35505	0.281

Table 2: Heston model. Posterior means and standard deviations over 10,000 iterations (top). CT = IF $\times$ N for the CPM scheme for $N = \beta\sqrt{T}$ and ρ selected such that $\kappa \approx 1.4$ at $\hat{\theta}$.

5.3 Linear Gaussian state-space model

We examine empirically the performance of the CPM method for multivariate state-space models using the particle filter with Hilbert sort described in Algorithm 2 and compare it to the PM method. Attention is restricted to a linear Gaussian state-space model which allows exact calculation of the likelihood and of the loglikelihood error $Z_T(\theta, U) = \log\{\widehat{p}(Y_{1:T} \mid \theta, U)/p(Y_{1:T} \mid \theta)\}$. Similar empirical results for non-linear non-Gaussian state-space models were observed.

We consider the model discussed in (Guarniero et al., 2017; Jacob et al., 2016) where $\{X_t; t \geq 1\}$ and $\{Y_t; t \geq 1\}$ are $\mathbb{R}^k$-valued with

$$X_1 \sim \mathcal{N}(0, I_k), \qquad X_{t+1} = A_\theta X_t + V_{t+1}, \qquad Y_t = X_t + W_t, \tag{55}$$

where $V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0_k, I_k), W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0_k, I_k)$ and $A_\theta^{i,j} = \theta^{|i-j|+1}$.

We use the transition density of $\{X_t; t \geq 1\}$ as proposal density within the particle filter. We first examine the achieved correlation between successive draws of $Z = \log\{\widehat{p}(y_{1:T} \mid \theta, U)/p(y_{1:T} \mid \theta)\}$ by running the CPM procedure holding the parameter fixed and equal to its true value $\theta = 0.4$. Next, we investigate the variance of $R = \log\{\widehat{p}(y_{1:T} \mid \theta', U')/p(y_{1:T} \mid \theta')\} - Z$ where $U' \sim K_\rho(U, \cdot)$ is the proposal when $\theta' = \theta$. This is performed for various values of T, with $N = \lceil \beta T^\alpha \rceil$ and $\rho = \exp(-\psi N/T)$ for $k \in \{2, 3, 4\}$.

We will now discuss the choice of α for state-space models. In sharp contrast to random effects models, we found empirically that there are dimension dependent limitations to the realized correlation that can be achieved through the particle filter with Hilbert sort. In particular we found that, due to resampling, the realized correlation is limited by $\min\{1 - c_1 N^{-1/k}, 1 - c_2\delta\}$ for some constants c_1, c_2, unless we set δ extremely small. Since the inefficiency tends to increase if we set δ too small, we balance the two terms by choosing $\delta = N^{-1/k}$, thus setting $\alpha = k/(k+1)$ for the following examples.

We run the CPM chain for 1000 iterations recording $\kappa^2 = \mathbb{V}(R)$ and $\sigma^2 = \mathbb{V}(Z)$. The values of β and ψ have been chosen so that they result in a particular target value of κ^2 as will be evident from the following tables. The asymptotic acceptance probability of the CPM scheme is thus in this case given by $\varrho_{\text{CPM}}(\kappa) := \varrho_{\text{U}}(\kappa) = 2\Phi(-\kappa/2)$ while it is $\varrho_{\text{PM}}(\sigma) = 2\Phi\left(-\sigma/\sqrt{2}\right)$ for the PM scheme (Doucet et al., 2015).

The results for $k = 2$ are reported in Table 3, where the two eigenvalues of A_θ are 0.56 and 0.24. The proposed scaling rule results in values of κ^2 which are approximately constant, remaining at values close to 2 for $T \geq 1600$. The implied acceptance probability of the CPM scheme $\varrho_{\text{CPM}}(\kappa)$ therefore settles at a value close to 0.5. By contrast, the marginal variance σ^2 increases at the expected rate $T^{1-\alpha}$ and accordingly the acceptance probability of the corresponding PM scheme, $\varrho_{\text{PM}}(\sigma)$, is very low even for $T = 100$. Similar results are found for the case $k = 3$, reported in Table 4, where the eigenvalues of A_θ are (0.6605,0.3360,0.2035) resulting in a model with moderately high persistence. In this case we set $\alpha = 3/4$. Although less dramatic, the implied gain of the CPM method over the PM method is substantial even for $T = 100$ and increases with T.

The full CPM procedure is now implemented for $T = 400$ and $T = 6400$ when $k = 2$ and $k = 3$ using the parameters of Tables 3 and 4. An autoregressive proposal is employed for θ which is based on the posterior mode and the second derivative at this point (Tran et al., 2016).

State dimension $k = 2$ with $\beta = 0.854$, $\psi = 0.12$, $\alpha = 2/3$						
T	N	$\delta = -\log \rho$	κ^2	σ^2	$\varrho_{\mathrm{CPM}}(\kappa)$	$\varrho_{\mathrm{PM}}(\sigma)$
100	18	0.0216	2.59	16.3	0.42	0.004
400	46	0.0138	2.71	20.5	0.41	0.0013
1600	116	0.0087	2.01	34.1	0.48	3.6×10^{-5}
6400	294	0.0055	2.07	49.7	0.47	6.0×10^{-7}
25600	742	0.0034	1.97	105.9	0.48	3.4×10^{-13}

Table 3: Linear state-space model. Results for $k = 2$ for varying T.

State dimension $k = 3$ with $\beta = 1.57$, $\psi = 0.042$, $\alpha = 3/4$						
T	N	$\delta = -\log \rho$	κ^2	σ^2	$\varrho_{\mathrm{CPM}}(\kappa)$	$\varrho_{\mathrm{PM}}(\sigma)$
100	49	0.0205	3.15	13.7	0.37	0.0089
400	140	0.0147	2.97	16.6	0.39	0.0039
1600	397	0.0104	3.44	26.7	0.35	0.00025
6400	1124	0.0074	3.03	34.1	0.38	3.66×10^{-5}
25600	3181	0.0052	2.69	49.4	0.41	6.74×10^{-7}

Table 4: Linear state-space model. Results for $k = 3$ for varying T.

The results for $k = 3$ and $T = 6400$ are shown in Figure 8. The mixing for θ is fairly rapid for the achieved value of $\kappa = 2.26$. The empirical distributions of Z under m and $\bar{\pi}$ are plotted (middle left) and are close to the theoretical distributions $\mathcal{N}\left(-\sigma^2/2, \sigma^2\right)$ and $\mathcal{N}\left(\sigma^2/2, \sigma^2\right)$ respectively, where $\sigma = 7.5$. The middle right plot and the third row show the draws of R, its empirical distribution and the associated correlogram arising from the CPM scheme. It is clear that R is approximately distributed according to $\mathcal{N}\left(-\kappa^2/2, \kappa^2\right)$, which is overlaid, but the correlogram decays slower than for random effect models and one-dimensional state-space models. The gain over the PM method is around σ^2 meaning we need around 50 times as many particles in the PM method to achieve similar results to the CPM scheme. When $T = 400$, we obtained $\kappa = 1.92$ and $\sigma = 4.30$ resulting in gains over the PM of approximately 18 fold. When $k = 2$, the gains are more impressive and are around 25 fold for $T = 400$ and 80 fold when $T = 6400$.

6 Discussion

The CPM method is an extension of the PM method using an estimator of the likelihood ratio appearing in its acceptance probability obtained by correlating estimators of its numerator and denominator. We have detailed implementations of this general idea for random effects and state-space models. For random effects models, we have provided theory to efficiently apply this methodology and have also verified empirically its efficacy for state-space models. In our examples, the computational gains over the PM method increase with T and can be over two orders of magnitude for large data sets. The CPM method is particularly useful for partially observed diffusions where sophisticated MCMC alternatives, such as particle Gibbs techniques, are inefficient.

From a theoretical point of view, in the random effects scenario, we have obtained a result suggesting that a necessary condition to ensure finiteness of the IACT of the CPM chain, as T increases, is to have N_T growing at least at rate $\sqrt{T}$. Our experimental results suggest that this condition is also sufficient and thus that the computational cost per iteration of the CPM method is $O(T^{\frac{3}{2}})$ versus $O(T^2)$ for the PM method. For state-space models, our empirical results indicate that this scaling degrades with the state dimension k and that we need N_T to grow at rate $T^{\frac{k}{k+1}}$ leading to a computational cost per iteration of order $O(T^{\frac{2k+1}{k+1}})$, up to a logarithmic factor[4], for the CPM method versus $O(T^2)$ for the PM method. It would be of interest but technically very involved to establish these results rigorously.

From a methodological point of view, it is possible in the state-space context to use alternatives to the Hilbert resampling sort to implement the CPM algorithm (Malik and Pitt, 2011; L'Ecuyer et al., 2018) and several such methods have been proposed following the first version of this work (arXiv:1511.04992); see, e.g., (Jacob et al., 2016; Sen et al., 2018). Empirical results in (Jacob et al., 2016; L'Ecuyer et al.,

[4]The particle filter with Hilbert sort has computational complexity $N_T \log N_T$ per observation.

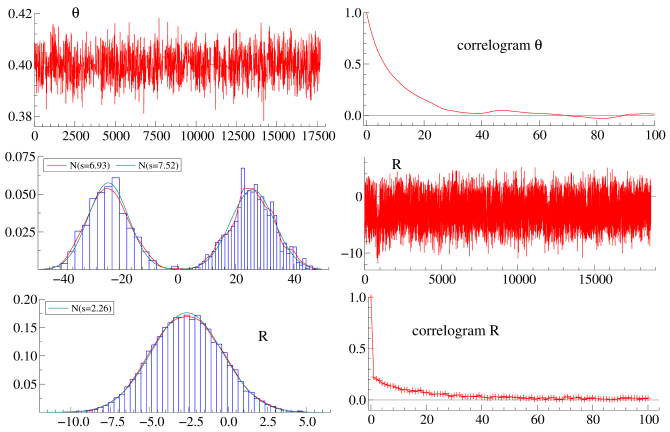

Figure 8: The CPM results for the 3-dimensional state space model with $T = 6400$. Top: parameter samples (left) and corresponding correlogram (right). Middle: Histograms of Z arising from m and $\bar{\pi}$ (left), draws of R (right). Bottom: Histogram of R (left) and correlogram (right).

2018) and our own experiments indicate that all these procedures provide roughly similar improvements over the PM method. One direction of interest is to use the sequential randomised Quasi Monte Carlo (QMC) algorithm, proposed and analysed by Gerber and Chopin (2015), within the CPM scheme by correlating the single uniform used to randomise the QMC grid. This is one motivation behind choosing the Hilbert sort procedure over alternative schemes, since this algorithm comes with theoretical guarantees. In a random effects context, the use of QMC has already been demonstrated to provide significant improvements (Tran et al., 2017). Finally, a sequential extension of the particle marginal MH algorithm (Andrieu et al., 2010), a PM method, has been proposed in (Chopin et al., 2013) and it would be interesting to develop an efficient sequential version of the CPM scheme.

Acknowledgments

The authors are grateful to the editor, associate editor, reviewers and Sebastian Schmon for their useful comments which have helped to improve the manuscript. Arnaud Doucet's research is partially supported by the Engineering and Physical Sciences Research Council, grant EP/K000276/1.

References

Andrieu, C., Doucet, A. and Holenstein, R. (2010) Particle Markov chain Monte Carlo methods (with discussion). *J. R. Statist. Soc.* B, **72**, 269–342.

Andrieu, C., Doucet, A. and Lee, A. (2012) Discussion of 'Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation' by Fearnhead, P. and Prangle, D. *J. R. Statist. Soc.* B, **72**, 451–452.

Andrieu, C. and Roberts G.O. (2009) The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, **37**, 697–725.

Andrieu, C. and Vihola, M. (2015) Convergence properties of pseudo-marginal Markov chain Monte Carlo algorithms. *Ann. Appl. Probab.*, **25**, 1030–1077.

Beaumont, M. (2003) Estimation of population growth or decline in genetically monitored populations. *Genetics*, **164**, 1139–1160.

Bérard, J., Del Moral, P. and Doucet, A. (2014) A lognormal central limit theorem for particle approximations of normalizing constants. *Electron. J. Probab.*, **19**, 1–28.

Berti, P., Pratelli, L. and Rigo, P. (2006) Almost sure weak convergence of random probability measures. *Stochastics*, **78**, 91–97.

Carpenter, J., P. Clifford and Fearnhead, P. (1999) Improved particle filter for nonlinear problems. *IEE Proc.*-F, **146**, 2–7.

Ceperley, D.M. and Dewing, M. (1999) The penalty method for random walks with uncertain energies. *J. Chem. Phys.*, **110**, 9812–9820.

Chopin, N. and Gerber, M. (2017) Sequential quasi-Monte Carlo: Introduction for non-experts, dimension reduction, application to partly observed diffusion processes. arXiv preprint arXiv:1706.05305, to appear in *Proceedings of Monte Carlo Quasi Monte Carlo 2016*.

Chopin, N., Jacob, P.E. and Papaspiliopoulos, O. (2013) SMC^2: an efficient algorithm for sequential analysis of state space models. *J. R. Statist. Soc.* B, **75**, 397–426.

Cotter, S.L., Roberts, G.O., Stuart, A.M. and White, D. (2013) MCMC methods for functions: modifying old algorithms to make them faster. *Statist. Sci.*, **28**, 424–446.

Crauel, H. (2003) *Random Probability Measures on Polish Spaces*. CRC Press.

Dalhin, J., Lindsten, F., Kronander, J. and Schön, T.B. (2015) Accelerating pseudo-marginal Metropolis-Hastings by correlating auxiliary variables. *Preprint arXiv:1511.05483*.

Del Moral, P. (2004) *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer-Verlag: New York.

Doornik, J.A. (2007) *Object-Oriented Matrix Programming Using Ox*, 3rd ed. London: Timerlake Consultants Press.

Doucet, A., Jacob, P.E. and Rubenthaler, S. (2013) Derivative-free estimation of the score vector and observed information matrix with applications to state-space models. *Preprint arXiv:1304:5768*.

Doucet, A., Pitt, M.K., Deligiannidis, G. and Kohn, R. (2015) Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika*, **102**, 295-313.

Flury, T. and Shephard, N. (2011) Bayesian inference based only on simulated likelihood: particle filter analysis of dynamic economic models. *Economet. Theor.*, **27**, 933–956.

Gentil, I. and Rémillard, B. (2008) Using systematic sampling selection for Monte Carlo solutions of Feynman-Kac equations. *Adv. Appl. Probab.*, **40**, 454-472.

Gerber, M. and Chopin, N. (2015) Sequential quasi Monte Carlo (with discussion) *J. R. Statist. Soc.* B, **77**, 509-579.

Geyer, C.J. (1992) Practical Markov chain Monte Carlo. *Statist. Sci.*, **7**, 473-483.

Gordon, N. J., Salmond, D. and Smith, A.F.M. (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proc.*-F, **140**, 107–113.

Guarniero, P., Johansen, A.M., and Lee, A. (2017) The iterated auxiliary particle filter. *J. Am. Statist. Ass.*, **112**, 1636–1647.

Heston, S.L. (1993) A closed-form solution for options with stochastic volatility with applications to bound and currency options. *Rev. Financ. Stud.*, **6**, 327–343.

Ionides, E.L., Breto, C. and King, A.A. (2006) Inference for nonlinear dynamical systems. *Proc. Natn. Acad. Sci. USA*, **103**, 18438–18443.

Jacob, P.E., Lindsten, F. and Schön, T.B.(2016) Coupling of particle filters. *Preprint arXiv:1606.01156*.

Johndrow, J.E., Smith, A., Pillai, N.S. and Dunson, D.B. (2016) Inefficiency of data augmentation for large sample imbalanced data. *Preprint arXiv:1605.05798*.

Koop, J.C.(1972) On the derivation of expected value and variance of ratios without the use of infinite series expansions. *Metrika*, **19**, 156–170.

L'Ecuyer, P., Munger, D., Lécot, C. and Tuffin, B.(2018) Sorting methods and convergence rates for array-RQMC: Some empirical comparisons. *Math. Comput. Simulat.*, **143**, 191–201.

Lee, A.(2008) Towards smooth particle filters for likelihood estimation with multivariate latent variables. M.Sc. Thesis, Department of Computer Science, University of British Columbia.

Lee, A. and Holmes, C.C.(2010) Discussion of 'Particle Markov chain Monte Carlo methods' by Andrieu, C., Doucet, A. and Holenstein, R. *J. R. Statist. Soc.* B, **72**, 327.

Lin, L., Liu, K.F., Sloan, J.(2000) A noisy Monte Carlo algorithm. *Phys. Rev.* D, **61**, 074505.

Lindsten, F. and Doucet, A.(2016) Pseudo-marginal Hamiltonian Monte Carlo. *Preprint arXiv:1607.02516*.

Lindsten, F., Jordan, M.I. and Schön, T.B. (2014) Particle Gibbs with ancestor sampling. *J. Mach. Learn. Res.*, **15**, 2145–2184.

Malik, S. and Pitt, M.K.(2011) Particle filters for continuous likelihood evaluation and maximisation. *J. Econometrics*, **165**, 190-209.

Murray, I. and Graham, M.M.(2016) Pseudo-marginal slice sampling. *Proc. 19th Conf. Artificial Intelligence and Statistics*, 911–919.

Nicholls, G.K., Fox, C. and Watt, A.M.(2012) Coupled MCMC with a randomized acceptance probability.

Preprint arXiv:1205.6857.

Papaspiliopoulos, O., Roberts, G.O. and Sköld, M.(2007) A general framework for the parametrization of hierarchical models. *Statist. Sci.*, **22**, 59–73.

Pitt, M.K., Silva, R., Giordani, P. and Kohn, R.(2012) On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *J. Econometrics*, **171**, 134–151.

Roberts, G.O., Gelman, A. and Gilks, W.R.(1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7**, 110–120.

Sen, D., Thiery, A.H. and Jasra, A.(2018) On coupling particle filter trajectories. *Statist. Comput.*, **28**, 461–475.

Sherlock, C., Thiery, A., Roberts, G.O. and Rosenthal, J.S.(2015) On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Statist.* **43**, 238–275.

Titsias, M.K. and Papaspiliopoulos, O.(2016) Auxiliary gradient-based sampling algorithms. *Preprint arXiv:1610.09641.* To appear in *J. R. Statist. Soc.* B.

Tran, M-N., Kohn, R., Quiroz, M. and Villani, M.(2017) The block pseudo-marginal sampler. *Preprint arXiv:1603.024845v4.*

Tran, M-N., Pitt, M.K. and Kohn, R.(2016) Adaptive Metropolis-Hastings sampling using reversible dependent mixture proposals. *Statist. Comput.*, **26**, 361–381.

van der Vaart, A.W.(2000) *Asymptotic Statistics.* Cambridge University Press.